Dunja Šešelja

Epistemic Evaluation in the Context of Pursuit and in the Argumentative Approach to Methodology

Proefschrift voorgedragen tot het bekomen van de graad van Doctor in de Wijsbegeerte
Promotors: Prof. Dr. Erik Weber en Prof. Dr. Joke Meheus

Ghent University

PhD Thesis:

Epistemic Evaluation in the Context of Pursuit and in the Argumentative Approach to Methodology

Author:

Dunja Šešelja

Supervisors:

Prof. Dr. Erik Weber,
Ghent University

Prof. Dr. Joke Meheus,
Ghent University

Reading Committee:

Prof. Dr. Tim De Mey,
Erasmus University Rotterdam, Ghent University

Prof. Dr. Teo Kuipers
University of Groningen

Prof. Dr. Thomas Nickles,
University of Nevada at Reno

Prof. Dr. Dagmar Provijn,
Ghent University

Prof. Dr. Maarten Van Dyck,
Ghent University

Acknowledgements

The research presented in this thesis has been carried out at the Centre for Logic and Philosophy of Science at the Department of Philosophy, Faculty of Arts and Philosophy, Ghent University. It has been supported by the Research Fund of Ghent University by means of the Research Project 01D03807, under the PhD grant by BOF (Bijzonder Onderzoeksfonds). I wish to thank Ghent University for allowing me to carry my research in an excellent working environment and under the working conditions any academic researcher can wish for.

In particular, I wish to express my deep gratitude to my supervisor Erik Weber for his continued support, invaluable suggestions, and careful guidance of my research. I am grateful to him for accepting me as his PhD student and helping me to clarify the main loci of my research, as well as for helping me to get acquainted with the literature and problems that have been of crucial relevance for my research topic. I am indebted to him for his willingness to co-author papers with me, by means of which I was able to learn a lot. This thesis would not have been possible without him. I am also deeply thankful to my co-supervisor Joke Meheus for her continued encouragement and support. Her comments and suggestions were especially helpful for the formal aspects of my research. Furthermore, I wish to thank Diderik Batens for all his support since my arrival to Belgium. He has greatly inspired my interest in formal approaches to philosophy of science and I have learned a lot from him.

I wish to thank to Christian Straßer who co-authored a number of papers with me, and whose willingness to have long and intensive philosophical discussions with me was invaluable support for my research. His sharpness in detecting problems and proposing paths for their possible solutions often led to a complete revision of the papers that constitute this thesis, and improved them significantly. I have learned a lot from him, and I hope our ongoing collaborations will result in many more joint projects.

I am deeply indebted to all the members of the Centre for Logic and Philosophy of Science for their warm welcome and constant support since my arrival to Ghent. Ever since I started my studies at Ghent University, I have felt at home, and in excellent working surrounding. I am thankful to many members of our Centre for helping me to broaden my philosophical horizon, sharpen my research skills, and stay aware of the many open questions that keep my curiosity and passion for philosophy awake.

I am also indebted to many members of our Department of Philosophy who have participated in various reading groups that greatly contributed to shaping and clarifying the ideas that constitute this thesis. I am also thankful to all the individual members of the Department with whom I have had many interesting philosophical discussions or who have given me valuable suggestions for my research.

I also wish to thank my professors, teachers, friends, and colleagues from Serbia who have encouraged my interest in philosophy and who have influenced my choice to work in this field. In particular, I wish to thank Mirko Aćimović,

who was my supervisor at the graduate studies at the University of Novi Sad, and who chose me to be his teaching assistant after my completion of the graduate studies. His encouragement of my interest in philosophy of science and of my decision to apply for a grant at Ghent University meant a lot to me. He would have been my PhD-supervisor if I had stayed at the University of Novi Sad. I am also indebted to Tomislav Kargačin and Tatjana Vukadinović who were my first philosophy teachers, and whose passion for philosophy greatly inspired me. I would not have taken the path of philosophy were it not for them.

Finally, I wish to thank my family for all the support they have given me, especially since my move to Belgium.

This thesis has been typeset in LaTeX in a GNU/Linux environment.
The image on the cover is *The Fall of Icarus*, 17th century, Musée Antoine Vivenel, taken from http://en.wikipedia.org/wiki/Icarus.
I am indebted to Christian Straßer for helping me with the layout of the thesis and LaTeX related issues.

Contents

Chapter

1 Introduction

This thesis is the result of my research on questions concerning theory change and certain aspects of rationality that underlies it. As the primary focus of my investigations emerged two issues: first, rationality governing the context of theory pursuit and second, formal models that can account for this type of rationality. In what follows I will clarify and motivate these two problem fields of my research.

1.1 Epistemic Justification and the Context of Pursuit[1]

1.1.1 Epistemic Justification of Scientific Theories

Science has many goals. Beside the practical ones, such as improving the life standards of citizens, or providing predictive control of our environment, there are also epistemic goals. Science should provide us with knowledge about the world. It should increase our understanding by providing explanations and accurate descriptions of natural or social phenomena.[2]

Epistemic justification is concerned with the latter type of goals. It is traditionally conceived of as providing standards for the acceptability of certain beliefs in the knowledge base or the cognitive system of an intelligent agent. Applied to a scientific theory, it provides criteria for its inclusion and acceptance into the grand corpus of our scientific knowledge. It concerns the question as to whether we have good reasons to consider it as being (approximately) truthful, empirically adequate, etc. Hence, epistemic justification is tightly connected to what Larry Laudan calls the context of acceptance, i.e. the context in which

[1] This section includes parts of my paper written together with Christian Straßer: (Šešelja and Straßer 201x), which is further presented in Chapter 3.

[2] Of course, these goals are interwoven. For instance, predictive power is also an epistemic goal.

scientists choose to accept a theory and thus treat it as if it were true (Laudan 1977, p. 108).

However, while it is a worthwhile epistemic goal to satisfy the criteria of theory acceptance, it is not the only one. A quick glance at the history of science reveals that scientific knowledge is highly dynamic and we shouldn't be all too assured with the theories we have accepted. Not just is it the case that theories often have to be altered and adjusted, but sometimes they have to be entirely replaced. Under sufficient pressure of anomalies we may be justified in no longer maintaining the belief that these theories are sufficiently good to be considered acceptable. These times of crisis we do not want to face empty-handed.

Therefore, another important epistemic goal of our scientific knowledge is achieving robustness with respect to these perturbations and conditions of uncertainty. If robustness is the ability to maintain performance in the face of perturbations and uncertainty (Stelling et al. 2004), then we can say that the scientific knowledge in a given domain is robust if it is able to maintain its key functions of explaining and helping us to understand the world, by means of avoiding and, if necessary, by overcoming scientific crisis. Clearly, the more robust our theories (in a certain domain) are with respect to these perturbations the more robust our scientific knowledge base as a whole (in this domain) is.[3] Although robust theories support this aim, since we cannot be sure that even the best theories withstand a possible future crisis we need more in order to ensure the robustness of the scientific knowledge as a whole. Recall that, as Otto Neurath famously remarked (Neurath 1932/1933 (1983, p. 92), scientists are "like sailors who have to rebuild their ship on the open sea, without ever being able to dismantle it in dry-dock and reconstruct it from the best components". Given the case that the old ship is about to sink and we cannot fix it anymore, we need to have (an)other "backup"-ship(s) available. Similarly, given the fact that even our best theories may fall in crisis, it is supportative of the aim of robustness to have alternative "backup"-theories around. These theories don't come from nowhere, but have to be thoroughly investigated and pursued.

This opens two perspectives on the composition and structure of scientific knowledge as a whole or in a give domain: (i) the flat perspective under which scientific knowledge is composed of accepted theories and (ii) the entrenched perspective under which scientific knowledge is composed and structured by layers of more and more entrenched theories. The degree of entrenchment may be measured by any standard of epistemic justification (such as for example the degree of coherence). At the most entrenched level we have the accepted theories. At the following levels we have alternative theories that may in times of crisis offer good backups for the accepted theories, or that may under further development eventually surpass the currently accepted theories. Although

[3]For an account of the robustness of theories (and/or their constitutive parts) see for instance (Wimsatt 2007, Calcott 2011) or (Chang 2004, p. 51-52). See also Footnote 5 in Chapter 3.

they do not (yet) suffice the criteria of actual epistemic justification for being accepted (e.g. they are not coherent enough) they are epistemically justified in a different sense since they, on the one hand, support the robustness of scientific knowledge, while, on the other hand, they are promising of developing into candidates for acceptance and in so far they serve the goal of adequacy and accuracy of scientific knowledge. We will thus say that a theory is *potentially epistemically justified* to the extent that it is promising of contributing to the epistemic goal of robustness and of developing into a candidate for acceptance.[4] In the following subsection we will suggest that a theory is epistemically worthy of further pursuit to the extent it is potentially epistemically justified.

1.1.2 The Context of Pursuit

According to Laudan, in addition to acceptance and rejection, pursuit and non-pursuit are the other two major cognitive stances that scientists can legitimately take towards research traditions (and their constituent theories) (Laudan 1977, p. 119). The notion of the context of pursuit resulted from the discussion on the traditional distinction between the context of discovery and the context of justification (proposed by Hans Reichenbach in the 30s' (Reichenbach 1938)) which, in view of many philosophers, needed to be refined by introducing an intermediate step. For example, Richard Tursman speaks of "the logic of pursuit and/or of preliminary evaluation of hypotheses", linking it to Charles S. Peirce's account of abduction as a logic of pursuit, according to which, there is a *prima facie* ground for pursuing a hypothesis which is capable of explaining certain surprising facts, which have been observed (Tursman 1987, p. 13-14). Imre Lakatos characterizes his "methodology of scientific research programmes" as consisting of "a negative heuristic", which tells us what paths of pursuit to avoid, and "a positive heuristic", which tells what paths to pursue (Lakatos 1978, p. 47). Ernan McMullin speaks of a "heuristic appraisal", which regards the research-potential of a theory (McMullin 1976). Thomas Nickles also discusses "heuristic appraisal" (Nickles 2006), as well as a "preliminary evaluation", "plausibility assessment" or "pursuit" as the context which "requires the comparative evaluation of problem-solving efficiency and promise, not simply the evaluation of completed research", in contrast to the traditional theories of confirmation (Nickles 1980, p. 21). Martin V. Curd argues that "not only is the logic of pursuit of more immediate practical relevance to scientific inquiry than the logic of probability but also that it is the only workable notion of a logic of discovery in the sense of a logic of prior assessment that one can formulate" (Curd 1980, p. 204). Finally, Laudan describes this intermediate step as "the context of pursuit" (Laudan 1977, 1980), and Laurie Anne Whitt as "theory promise" or "theory pursuit" (Whitt 1990, 1992).

Nevertheless, the question of pursuit has often been left out of the accounts of epistemic justification. Even though some of the above mentioned authors

[4] Also Sven Ove Hansson (2003) makes –in reference to David Makinson– the distinction between actual and potential justification of beliefs.

discuss the nature of the context of pursuit or the possible logic of pursuit, and even though it has often been pointed out that such a prior assessment already embraces elements of justification (Schickore and Steinle 2006a, p. viii), there has been little to no consideration of this question in the concrete accounts of epistemic justification.[5] In contrary, pursuit worthiness has been mainly discussed in view of an interwoven set of epistemic and non-epistemic values, the latter referring to social, ethical, or political values or personal interests of scientists (e.g. (Nickles 2006), (Kitcher 2001, Chapter 9), (Douglas 2009, Chapter 5)).

As we have suggested in the previous subsection, pursuit worthiness is a valid subject of epistemic justification that needs to be addressed in a different way than theory acceptance. Scientific theories clearly do not suddenly come into existence complete and fully equipped with an explanatory apparatus that would satisfy the standards of theory acceptance. Their origin lies in ideas and hypotheses that have been thoroughly investigated, reformulated, corrected. But at the same time, young theories can be promising of developing into good backups for the currently established theories, and eventually even into acceptable ones. Hence, from an epistemic perspective, what we are concerned with in the context of pursuit is not the question as to whether a theory is acceptable, but as to whether there are good epistemic reasons for its further pursuit. We will say that a theory is *epistemically worthy of pursuit* to the extent that it can be shown to have a promising potential for contributing to those epistemic goals that determine theory acceptance, as well as to the value of robustness. In other words, a theory is epistemically worthy of pursuit to the same degree that it is potentially epistemically justified.

To be sure, the evaluation in the context of pursuit, as a part of the scientific practice, is certainly not exclusively epistemic. Many other non-epistemic factors play a role in deciding which problems to tackle, which methodology is ethically acceptable, etc. But this does not imply that epistemic values do not have a place in such an evaluation. In contrary, debates among scientists about the further pursuit of emerging scientific theories are often focused on novel explanations and predictions that the given theory offers, its consistency and compatibility with theories from other scientific domains, etc. Having good epistemic reasons for the further investigation of a theory is an important criterion for deciding about its pursuit, though it is not the only one that is practically relevant.

It is also important to notice that the question of epistemic pursuit worthiness is different from the question: Which theory should an individual scientist actually pursue? Showing that a theory is epistemically worthy of pursuit does not imply that *each* scientist should engage in its actual pursuit, since more

[5] The closest to an epistemic treatment of pursuit worthiness came Laurie Anne Whitt who, in response to McMullin's approach, remarked that "There seems to be no reason to accept the stipulation that epistemic appraisals are limited to contexts of acceptance." (Whitt 1992, p. 616). In addition, Chang's coherentist epistemic iteration (Chang 2004) addresses some aspects of the context of pursuit even though he does not explicitly discuss the notion of pursuit or pursuit worthiness.

than one theory can simultaneously be epistemically worthy of pursuit. The fact that some of them are epistemically worthy of pursuit should not be confused with specific preferences individual scientists may have when choosing which of these theories to work on. A rational division of labor in a given scientific discipline depends on the epistemic status of all available theoretical candidates, as well as on some non-epistemic factors, such as the number of scientists working in the field, the financial resources, personal interests and expertise of the scientists etc. Therefore, an account of the epistemic pursuit worthiness provides tools that (together with some other elements) play a central role in determining the rational division of labor in a given scientific community.

In the first part of this thesis I discuss some of the central questions raised by the neglect of the context of pursuit in philosophy of science and epistemology. First of all, in Chapter 2 I disambiguate between different notions of pursuit and pursuit worthiness. Epistemic pursuit worthiness will be explicated as one of the notions that plays a key role in the evaluation of scientific theories. In Chapter 3 I present a coherentist account of epistemic justification suitable for the context of pursuit, that I have developed together with Christian Straßer. In Chapter 4 I apply this framework to the revolution in geology in order to investigate the pursuit worthiness of Alfred Wegener's theory of continental drift. This chapter has resulted from a joint work with Erik Weber.

1.2 Formal Modeling – Argumentation Frameworks and Adaptive Logics

1.2.1 Epistemic Evaluation from an Argumentative Point of View

In the last couple of decades discussions in the field of philosophy of science and scientific methodology have witnessed a growing conviction that a rule-based algorithmic approach to theory appraisal is problematic (e.g. (McMullin 1982, p. 17), (McMullin 1984, p. 56), (Kuhn 1996, p. 198-199), (Kuhn 2000, p. 200)). One possible attempt to preserve the normative idea of rationality in spite of abandoning the idea of a static, universally applicable scientific method can be found in more rhetorically minded approaches to scientific reasoning, such as Marcello Pera's (1994) or Marcelo Dascal's (2000). Instead of an algorithmic assessment of scientific theories, Pera and Dascal emphasize the evaluation in view of the argumentative context underlying the given episode in the history of science. While formal approaches to scientific reasoning have been mainly focused on the logical form of arguments (that is, the nature of the inference relation), both Pera and Dascal show that scientific debates (Dascal's *controversies*) are typically not resolved by derivational reasoning that is characteristic for logic but rather by scientists exchanging arguments and trying to convince each other by giving reasons that substantiate their points.

In view of this argumentative shift in methodology, an account of epistemic justification, suitable for either the context of pursuit or the context of acceptance, has to allow for its constitutive criteria to be determined in view of scientific debates relevant for the given historical context. That means that the nature of the criteria, or their specific preference order cannot be pre-defined, but it is dependent on the specific historical context. The specificities of historical contexts can be found by a close analysis of scientific debates in the given domain at the time.

Chapter 5 starts from these insights and reexamines the previously presented account of epistemic justification from such an argumentative perspective. We will take a closer look at Kuhn's views towards theory evaluation in the context of pursuit, as well as at his notion of persuasion. The latter notion will point to the argumentative approach to scientific methodology, which will be presented in view of McMullin's and Pera's views.

Let us also mention that a discussion on the relation between Kuhn's views and coherentist epistemology is presented in the Appendix of this thesis. Even though related to the above points, this discussion is written as a reply to (Kuukkanen 2007) and hence, stands on its own. One of the points we make in this paper regards Kuhn's stance on the rationality underlying interparadigmatic changes and its relation to coherentist epistemology. The paper is a result of a joint work with Christian Straßer.

1.2.2 Explanatory Argumentation Frameworks

The argumentative approach to methodology motivates an argumentative approach to formal modeling of theory assessment as well. Formal theories of argumentation have been extensively researched within the fields of artificial intelligence, philosophy, logic and computer science. One of the most influential formal accounts of argumentation is Phan Minh Dung's abstract argumentation framework (Dung 1993, 1995). The significance of Dung's approach derives from the fact that it abstracts away from the nature of arguments and argumentation rules, which allows us to focus on the interplay of arguments rather than on their specific structure. The fruitfulness of this framework stems also from the fact that it is easily enhanceable with additional properties and the fact that it is useful in different application contexts. However, the abstract argumentation framework has so far not been applied to the modeling of theory choice and scientific debates. Moreover, the framework has not been enhanced with explanatory capabilities, which play a significant role in scientific reasoning. Chapter 6 presents this enhancement, which we call the Explanatory Argumentation Framework, and its application to the modeling of scientific debates. We will show that such a framework allows for a comparison of different sets of arguments in view of their explanatory virtues. Furthermore, we will offer a set of criteria which are useful for the demarcation of rivaling scientific views in terms of arguments, as well as for an evaluation of such views in terms of their argumentative and explanatory properties. We will show that such a modeling is suitable for theory evaluation in both the context of pursuit

and the context of acceptance. This chapter is a result of a joint work with Christian Straßer.

1.2.3 Adaptive Logic Framework for Abstract Argumentation

The field of abstract argumentation opens the question how to represent the reasoning underlying selection procedures of acceptable sets of arguments in a formal logical way. According to Dung, "Logic-based knowledge bases can be viewed as argumentation systems where the knowledge base is coded in the structure of the arguments and *the logic* is used to determine the acceptability of the arguments." (Dung 1995, p. 856, italics added). By representing the acceptability of arguments in a proof theoretic manner, an additional approach to formalizing scientific reasoning can be offered. Such an approach differs from logical formalizations of inferences made in scientific reasoning (such as induction, abduction, etc.) in view of the given propositions. As John L. Pollock remarks: "Constructing arguments is one thing. Deciding which conclusions to accept is another. ... The conclusions that ought to be believed are those that are undefeated." (Pollock 1987, p. 7). Hence, in this type of logical systems, the focus is not on the derivation of propositions in view of other individual propositions, but on the derivation of arguments, and thus of the conclusions made by them, that are considered acceptable in view of the whole set of arguments constituting the given debate.

The adaptive logics have shown to be a fruitful system for this purpose. In Chapter 7 I present an adaptive logic framework that provides a proof theory for all the standard selection procedures of abstract argumentation frameworks. This part has resulted from a joint work with Christian Straßer.

Chapter

2 Diambiguating the Notions of Pursuit and Pursuit Worthiness

✎ *I wish to thank Christian Straßer and Erik Weber for valuable comments on a former version of this chapter.*

Summary In this chapter we will disambiguate between different types of pursuit which belong to the scientific practice, and between different types of pursuit worthiness, that is, different ways in which a theory can be worthy of pursuit. With respect to the notion of pursuit, we will distinguish between, on the one hand, the pursuit of explanatory theories and, on the other hand, the pursuit of scientifically relevant phenomena, entities, technological developments, etc. With respect to the notion of pursuit worthiness, we will distinguish, first, between the epistemic and non-epistemic kind of pursuit worthiness, and second, between the pursuit worthiness regarding individual rationality and the one regarding group rationality. Finally, in view of these distinctions we will explicate the notions of pursuit and pursuit worthiness that will be of our primary interest in this thesis.

2.1 Introduction

To pursue a theory means to engage in its further investigation, aiming at its development or a development of its variants. Whether a theory is worthy of pursuit is usually assessed in terms of epistemic as well as non-epistemic concerns. As we have announced, our primary interest in this thesis is the epistemic pursuit worthiness of scientific theories. But before we present our framework suitable for the evaluation of epistemic pursuit worthiness, we should clarify the idea of pursuit and pursuit worthiness. More precisely, we should explicate in which way pursuit of scientific theories differs from other types of pursuit that are of scientific interest, and in which way epistemic pursuit worthiness differs from the notion that involves both epistemic and non-epistemic criteria.

Finally, the distinction between individual and group rationality underlying the context of pursuit should be explicated as well.

Some of these aspects of pursuit are also characteristic of a pursuit conducted by a detective who is investigating a criminal case. Thus, before we turn to the scientific enterprise, let us take a look at how a detective's pursuit is evaluated. This might help us in noticing that similar aspects can be distinguished in the case of scientific pursuit as well.

First, if a detective is to investigate a new murder case, she will try to build a hypothesis of how the murder occurred. There are different ways in which it can be evaluated whether her hypothesis is worthy of pursuit. On the one hand, our detective may need to show (for example, when presenting the case to her super-ordinate) why her hypothesis is *epistemically* worthy of pursuit. This includes showing that there is evidence which indicates that her hypothesis best explains the given case. Moreover, she is to show how she plans on proceeding with her investigation. That includes showing that her pursuit will be based on the heuristics that can allow for further evidence to be collected and used to corroborate or falsify the initial hypothesis. For example, in order to investigate the main suspect, the detective might propose introducing wiretaps on the suspect's phone line. However, in spite of being epistemically interesting, such a pursuit might conflict with certain ethical or political concerns. For example, her proposal to implement wiretaps might affect the privacy of people who are not directly involved in the given case. Or her pursuit could lead to the evidence that certain politicians, who are directly responsible for the funding of the police department, are involved in the case. In the former case, the pursuit might be assessed as *ethically* problematic in its current form. In the latter case, the pursuit might be assessed as *politically* problematic and moreover, financially unfeasible. In other words, in spite of being epistemically worthy of pursuit, the investigation does not have to be *practically* worthy of pursuit, where by practical pursuit worthiness we mean the assessment of the given pursuit in view of both epistemic and non-epistemic values that are determined by the given socio-political framework. Finally, her case could be both epistemically and practically worthy of pursuit, but, for instance, due to certain reductions in the budget of the police department, not actually pursued.

Second, we can imagine that another detective, working on a different case, obtains through his own investigation some new evidence relevant for the above mentioned murder as well. Moreover, he might develop a different hypothesis for who the main suspect in this murder is. We can imagine that even though his hypothesis is supported by some evidence, it cannot account for all the evidence provided by his colleague, while her hypothesis cannot account for some of his evidence either. Hence, both hypotheses could turn out to be epistemically worthy of pursuit, that is, in view of all the available evidence. Note, however, that this does not mean that investigative plans for both of them would necessarily be equally worthy of pursuit in other respects as well: they could differ in ethical, political or social aspects. But let us assume that

they are in these respects unproblematic.[1] In this case, it would be rational for both detectives to further pursue their respective hypotheses, until they obtain some more decisive evidence.

This example nicely shows different aspects of pursuit in general, and different notions of pursuit worthiness. We have first seen that the fact that a hypothesis and an investigation based on it are epistemically worthy of pursuit may not coincide with its practical pursuit worthiness. Second, more than one hypotheses may simultaneously be practically worthy of pursuit. Hence, the fact that one of them is assessed as such does not mean that each detective working on the case should engage in its actual pursuit. Different detectives may simultaneously conduct different investigations regarding the same case. In the remainder of this chapter we will show that these distinctions apply in a similar way to the notions of pursuit and pursuit worthiness regarding scientific theories.

2.2 The Pursuit Worthiness of Theories and the Pursuit Worthiness of Other Types of Scientifically Relevant Questions

Our detective case showed certain aspects of pursuit that are similar to the pursuit of scientific *theories*. But before we take a look at different types of pursuit worthiness regarding scientific theories, we should first distinguish the pursuit of scientific theories from other types of pursuit, that may also be scientifically relevant.

Depending on what kind of issue is being investigated, we can distinguish between a pursuit of scientific theories and other types of pursuit, such as those regarding scientifically relevant phenomena, entities or technological developments.[2] These types of pursuit are often interwoven within the same research. For instance, the pursuit of Wegener's theory of the continental drift, which was explanatory of different geological explananda, implied the pursuit of the phenomenon of drift itself. Therefore the pursuit worthiness of the theory of continental drift implied that the phenomenon of drift was worthy of pursuit as well.

However, the criteria used for the evaluation of the pursuit worthiness of explanatory theories may not always be suitable for the evaluation of the former type of pursuit.[3] A scientifically relevant phenomenon worthy of pursuit

[1] Of course, what counts as problematic here depends on the ethical and political standards one adopts. See also the discussion in Section 2.3.

[2] Martin Carrier makes a similar distinction between "epistemic research" as the search for understanding, characterized by knowledge-guided mode of problem selection, and "application-oriented research" as the search for utility (Carrier 2010). However, our distinction is different since a scientifically relevant phenomenon does not necessarily belong to a search for utility (see the example of the continental drift bellow).

[3] Even though we are placing here very different types of pursuit (that of phenomena, entities, technological inventions) in the same group, this is only due to the fact that in this thesis we will be interested in pursuit and pursuit worthiness of explanatory theories. Hence

may be, for example, a certain statistical correlation. For instance, pursuing correlations such as those between smoking and lung cancer can be worthy in view of certain epistemic and social reasons. However, once we have shown that the correlation holds, we are also interested in a theory that explains it. Whether such a theory is worthy of pursuit or not needs to be evaluated in a different way (for instance, by taking a look at how good explanations the theory offers, how well connected it is with other scientific theories, etc.).

Another example of the pursuit of scientifically relevant phenomena is the investigation of the question as to whether there is extra-terrestrial life, as it has been done by various SETI (the Search for Extraterrestrial Intelligence) projects. For this investigation to be worthy of pursuit, we need to show that there is a methodology that provides the heuristic of our investigation, that there is a certain level of likelihood of the success in finding the extraterrestrial intelligent life forms, as well as that the overall epistemic and non-epistemic benefits of such an investigation outweigh the possible dangers (see (Kukla 2001)). In contrast, a theory that aims at explaining the extraterrestrial life can be worthy of pursuit only after the evidence of such life forms has been found. Moreover, when evaluating the pursuit worthiness of such a theory, we will not only be interested in how good its heuristic is, but also in how good its current explanations are, and why the theory seems promising of developing into a highly explanatory one. Even though some criteria of pursuit worthiness may overlap, some others will be different.

Yet another example of pursuit that should be distinguished from a pursuit of explanatory theories is the one regarding technological developments. Pursuing the invention of an instrument, apparatus, machine, etc. could be a part of a pursuit of a certain explanatory theory. For instance, a pursuit of nuclear weapons can be seen as a part of the pursuit of theories within the domain of nuclear physics, where the former was not only an application of the latter, but it also served to produce additional evidence for it. Clearly, there are good reasons why a pursuit of such a technology may be considered highly ethically problematic and in so far unworthy of being conducted. But this does not mean a pursuit of theories in the domain of nuclear physics is unworthy as well, in case they offer alternative ways of obtaining the evidence regarding their hypotheses. When we evaluate whether a given technological development is worthy of pursuit, we are interested in how useful such a technology could be, how easy it would be to handle it, what the benefits and dangers of such a pursuit are, etc. In contrast to scientific theories, technological developments do not need to aim at offering scientific explanations (though they may indeed make use of scientific explanations, that serve as guidelines in the construction of the given technology).

To sum up: on the one hand, the pursuit of phenomena, entities, and technological developments, and on the other hand, the pursuit of explanatory theories belong to different types of pursuit, which may be tightly connected.

---

the aim of the above distinction is to delineate the latter type of pursuit, rather than to properly define the former.

Nevertheless, evaluating their respective pursuit worthiness may require different criteria of evaluation. In the following two sections we will take a closer look at the latter type – pursuit of theories.

2.3 Epistemic and Non-epistemic *viz.* Practical Notions of Pursuit Worthiness

Let us now get back to our example of a detective case. The first type of distinction that we have seen there is the one regarding the question: According to which criteria is pursuit worthiness of a theory evaluated?

As we have already mentioned, it has been common in the philosophical literature to relate the epistemic justification primarily to the context of acceptance. In contrast, theory evaluation in the context of pursuit has usually been related to a joint set of epistemic and non-epistemic criteria. For example, Thomas Nickles makes a distinction between *epistemic appraisal (EA)* and *heuristic appraisal (HA)*:

> EA attends to truth-conductive features of justification and decision-making, while HA attends to a variety of heuristic and pragmatic considerations relating to economy of research. ... HA evaluates the promise or potential fertility and feasibility of further work on a problem, research program, theory, hypothesis, model, or technique. (Nickles 2006, p. 159)

Nickles argues that HA cannot be reduced to EA, nor can it be described as its derivative (p. 164). Moreover, according to him HA is primarily concerned with pragmatic aspects of the scientific research, namely, its fertility and its practical realizability (p. 165). It regards questions such as: "Where do we go from here? What would be a good project to do next?", "Is the project feasible for anyone right now? For us?" (p. 167). Nickles thus regards HA as an evaluation which is done in view of not only epistemic criteria, but also – and essentially – in view of non-epistemic criteria. When performing HA, scientists must take into account "external factors" such as those regarding the question as to "whether their research is likely to be funded, whether the lab director or department head will look favourably upon this project; whether enough laboratory space, equipment, and expert technical assistance is available" etc. (p. 169).

A similar approach to the notion of pursuit can be found in Philip Kitcher's 2001.[4] Kitcher proposes a detailed account of how a scientific inquiry should be organized, where the notion of pursuit is understood in terms of both epistemic and non-epistemic standards (Kitcher 2001, Chapter 9). However, he recognizes a possible conflict between epistemic and non-epistemic pursuit worthiness. Arguing that restrictions on a free inquiry are sometimes justified, he writes:

[4]Another example would be Heather Douglas' discussion of epistemic and non-epistemic values that jointly play a role in assessing pursuit worthiness of research processes (Douglas 2009, Chapter 5).

> Respecting rights comes at a price, and it's important that the price be distributed fairly. In situations where free inquiry would unfairly increase the burden on those who are already disadvantaged, there can be no right to free inquiry. (Ibid., p. 103)

Without going into a discussion on Kitcher's view on free inquiry, there is an important point about pursuit that Kitcher raises here. The conflicting interests regarding pursuit can be presented in terms of different *prices* that need to be payed if the pursuit is conducted. For instance, if the research involves a certain ethically problematic methodology, we may say that the *ethical* price to pay is too high, and hence, the pursuit in this form should be avoided. Similarly, we may ask whether the *epistemic* price to pay is too high if a certain pursuit (that is, a scientific project involving it) is no longer financed. The epistemic price could, for instance, refer to the abandoning of this research altogether, or to giving up on certain aspects of it, which would, if realized, result in an epistemic benefit. Those arguing for the rejection of a further financial support may say that the *economic* price to pay outweighs the epistemic benefit in case the theory is pursued (for example, due to the fact that the theory is epistemically not very interesting, while its pursuit requires a huge investment). Of course, how one weighs out different "prices to be paid" is also dependent on social and political interests that determine what means that some factors outweigh the others.

Altogether, what this tells us is that a decision regarding the undertaking of a given pursuit involves the weighting of different epistemic and non-epistemic factors, where we estimate which epistemic, ethical, economic, political, etc. prices to be payed are right. Deciding whether a scientific theory is worthy of pursuit in view of all these respects amounts to, what we shall call, *practical pursuit worthiness*. Nevertheless, we can also speak of *epistemic pursuit worthiness* alone. This notion concerns the question as to whether a theory is worthy of pursuit in view of epistemic values, disregarding whether the non-epistemic factors have been satisfied.

The following example should help us to clarify some more aspects of the relation between the epistemic and the practical pursuit worthiness. Imagine that a further investigation of a developing scientific theory is, in principle, technically realizable but it conflicts with certain ideological views of the government and with the institutions responsible for funding of scientific research. When evaluating such a research proposal according to the criteria comprising of both epistemic and non-epistemic values, the scientists would have to conclude that their project is simply not worthy of pursuit, since it is politically controversial, and is thus unlikely to be financed. Nevertheless, from an epistemic point of view, their project may be very well worthy of pursuit. Moreover, it may be practically worthy of pursuit from a perspective rooted in a different social-political values.

There are two points that can be made in view of this example. First, depending on the ethical and political framework we adopt, we can have different standards of the practical pursuit worthiness. On the one hand, we can

root the notion of pursuit worthiness in a normative framework which is *external* to the *status quo*, and which may thus differ from it. Kitcher's approach, based on democratic values, is an example of this notion. According to his account of (what we have called) practical pursuit worthiness, a theory would not necessarily be practically unworthy of pursuit if it conflicted with some ideological views of the given government. Moreover, in such a framework, if a theory is practically worthy of pursuit, then it is also epistemically worthy of pursuit (though not be the other way around). On the other hand, the notion of pursuit worthiness can be built in view of the normative framework that is *internal* to the *status quo*. In our example from above, the given theory is not practically worthy of pursuit in this sense of the term.

Second, an assessment of epistemic pursuit worthiness may give a different result than an assessment of practical pursuit worthiness. The usual way to define epistemic values[5] is as those that are considered to be conducive of the cognitive goals of science, such as truth, problem-solving efficiency, etc.[6] Different scholars have offered different lists of values that fall in this category, especially with regard to the problem of theory choice. Kuhn, for instance, lists accuracy, consistency, scope, simplicity, and fruitfulness as "the standard criteria for evaluating the adequacy of a theory" (Kuhn 1977, p. 322). According to McMullin, epistemic values include predictive accuracy, internal coherence, external consistency, unifying power, fertility and simplicity (McMullin 1982, p. 15-16). Indeed, different approaches to the problem of theory choice (given in view of different methodological, epistemological, and other philosophical views) give different lists of epistemic criteria relevant for the evaluation of scientific theories.

However, when it comes to the epistemic pursuit worthiness it is important to mention that in addition to the usual epistemic standards[7] it also includes certain *pragmatic* elements. Recall that in our example of a developing scientific theory we have assumed that its further investigation was, in principle, technically realizable. *Technical realizability* is a criterion that obviously introduces a pragmatic element. This criterion concerns the feasibility of a research and its methodological requirements in view of the current technological achievements, and it is directly linked to the heuristics of the given research. It concerns the question as to whether the heuristics of the theory allows for further evidence

[5]We shall use the terms "value", "criterion", and "standard" in this context interchangeably.

[6]Ernan McMullin, for example, characterizes them as values that are "presumed to promote the truth-like character of science, its character as the most secure knowledge available to us of the world we seek to understand. An epistemic value is one we have reason to believe will, if pursued, help toward the attainment of such knowledge." (McMullin 1982, p. 18). Similarly, Phyllis Rooney defines them as "those [values] that are usually taken as constitutive of the knowledge and truth-seeking goals of the enterprise of science" (Rooney 1992, p. 15).

[7]Even though we speak here of the usual set of epistemic criteria, it is important to mention that in order to be suitable for theory evaluation in the context of pursuit, they need to be formulated in a different way than those suitable in the context of acceptance (see Chapter 3).

to be collected and used to support the given hypotheses or to confront them with possible anomalies. Of course, in case the required technology itself first needs to be pursued, the epistemic pursuit worthiness of the given theory will directly depend on the epistemic pursuit worthiness of the given technology.[8]

But why is making the distinction between epistemic and practical pursuit worthiness important? In which way does it help us to understand the context of pursuit in contrast to the above mentioned accounts of pursuit that already include both epistemic and non-epistemic factors relevant for assessing pursuit worthiness? The main significance of the distinction is that it allows us to focus only on epistemic properties of a theory. On the one hand, this is important for the evaluation of certain episodes from the history of science and the rationality underlying views of scientists with regard to the pursuit of certain theories. For example, if a pursuit of a theory was rejected due to some epistemic reasons, we will be interested in assessing whether such reasons were justified or whether the theory was, in fact, epistemically worthy of pursuit. On the other hand, the epistemic approach may help us in clarifying the reasons why a certain pursuit is favored. For instance, it could be found out that a pursuit of a certain theory is supported because of certain political reasons (and thus considered as practically worthy of pursuit in the second sense that we have explicated) in spite of being epistemically not very attractive.

As we have already pointed out, showing that a theory is epistemically worthy of pursuit does not mean that it should actually be pursued, since it may be problematic in some other respects that fall under practical pursuit worthiness. But even if a theory is practically worthy of pursuit (in either of the above two senses), that does not mean it should necessarily be actually pursued. If certain conditions, independent of the theories themselves but required for their pursuit, are not fulfilled, their pursuit might be difficult to realize. For instance, the budget from which scientific projects are financed may temporarily be too small to support all the candidates that are practically worthy of pursuit. Or the number of scientists working in the given domain may be too small to cover all the theories that are practically worthy of pursuit.

[8]Furthermore, *how* epistemic criteria are evaluated may depend on some non-epistemic factors. For example, if we want to assess whether a theory has exhibited certain growth, which would help us in judging whether it has remained worthy of pursuit, we will have to take into account the number and the expertise of scientists working on the theory or the appropriate funding that allows for the required resources (see Chapter 3). That does not mean that the epistemic evaluation is not *epistemic* in character (we are still evaluating the *epistemic* growth of the theory), it just means that our expectations regarding the epistemic standards are in this sense context dependent. In other words, the non-epistemic factors determine in which respects the conditions for fulfilling the epistemic standards have been met, and what can thus be expected from the given theory.

2.4 Pursuit Worthiness Regarding Group and Individual Rationality

Similarly to our detective case where we have seen that more than one hypothesis may be worthy of pursuit, when it comes to the scientific enterprise, more than one scientific theory may be simultaneously worthy of pursuit as well. This holds for both epistemic and practical notions of pursuit worthiness. But when we say that more than one theory may be simultaneously worthy of pursuit, what we mean is worthy of pursuit for the given *scientific community*, that is, for the scientists working in the given scientific domain. Hence, such a notion of pursuit worthiness regards rationality of a group, rather than rationality underlying decisions of an individual scientist. If a theory is worthy of pursuit for the given scientific community, that does not mean that each individual scientist working in the given domain is supposed to actually pursue it (see also (Whitt 1990, p. 476-477)).

So far we have used the term "pursuit worthiness" only to refer to the communal notion, which will also be our primary concern in this thesis. Nevertheless, it is worth noticing that the individual pursuit worthiness is usually used as the practical pursuit worthiness (that is, it can be considered as a subtype of the practical pursuit worthiness). What is worthy of pursuit for an individual scientist depends not only on epistemic criteria, but also on different non-epistemic factors, such as the expertize of the scientist in a particular domain, his or her personal interests and preferences, etc. (see also (Whitt 1990, p. 479)).

Failing to recognize the distinction between the *individual* and the *communal* pursuit worthiness can lead to ambiguities in understanding the notion of pursuit. For instance, when Nickles writes: "Deciding that a defective theory or model is worthy of further pursuit amounts to launching or continuing a research program" (Nickles 2006, p. 168) – we can agree with this statement only if we link both phrases "worthy of pursuit" and "launching or continuing a research program" either to a group or to an individual scientist. On the one hand, if a theory is worthy of pursuit for the given scientific community, then *some scientists* (and in that sense, the scientific community) working in this field should actually pursue it. On the other hand, if a theory is worthy of pursuit for a certain scientist, then she or he should pursue the theory. However, the sentence would get a completely different meaning (which would conflict with our view from above) if we interpreted it as saying that if a theory is worthy of pursuit for an individual scientist, then all other scientists should engage in its pursuit, or as saying that if a theory is worthy of pursuit for the community, every individual scientist should actually pursue it.

An important consequence of the fact that more than one theory may at the community level be simultaneously worthy of pursuit[9] is that we do not

[9]Of course, it could also be the case that more than one theory is at the same time worthy of pursuit for an individual scientist, who then engages in the actual pursuit of the given theories.

necessarily need a dissensus among scientists regarding the pursuit worthiness of theories in order to have a diversity in the context of pursuit. Scientists may *agree* that different theories are worthy of pursuit, and yet, each of them may engage in a pursuit of only one of them.

Surprisingly, this quite obvious point has been overseen by a number of authors who have discussed theory choice, especially with regard to scientific controversies. In order to have the diversity of theories that are actively pursued, they have often pointed out a disagreement among scientists as crucial for this purpose. We will take a look at three such examples.

Our first example is from (Hoyningen-Huene 2006). In his discussion of Kuhn's stance on theory choice Paul Hoyningen-Huene reminds the reader that, according to Kuhn, even though a given scientific community may have a set of shared values of theory choice, each individual member of the community will specify these values with respect to their content and their mutual weight (p. 127). In view of this he explains that the consensus can be reached only if, in spite of these "individual value differences", an agreement among scientists about which theory should be accepted eventually emerges. In order to show why such a decision procedure is justified, Hoyningen-Huene suggests the following:

> ...it must be shown that also the individual value differences that lead to disagreement in the phase of extraordinary science but disappear from the result of communal theory choice, are rational means towards the cognitive goals of science. The main idea here is that these differences make a rational disagreement during the phase of extraordinary science possible. *This disagreement is vital for the distribution of risk in a situation of epistemic uncertainty as no one knows, which candidate for paradigmatic theory will be successful.* (p. 128, italics added)

Without going into discussion on Kuhn's view on theory choice,[10] let us notice the following claim in the above quote: that the disagreement among scientists is *vital* for the diversity of pursued theories. Nevertheless, while a disagreement among scientists regarding the question, which theories are worthy of pursuit, may be *fruitful* in this sense, it is not clear why it would be vital for it. If scientists can evaluate more than one theory as worthy of pursuit in the communal sense of the term, such an *agreement* would allow for the diversity as well. The division of labor based on their individual preferences may not include a disagreement either: a scientist may find more than one theory worthy of pursuit for herself, but decide to pursue only one, while agreeing that her colleague should pursue the other.

Our second example is from (Rueger 1996), which shows a similar neglect of this point. With regard to the pursuit of scientific theories, Alexander Rueger writes:

[10] Indeed, the emphasis on rational disagreement is present in Kuhn's work: see, for example, (Kuhn 1977, p. 220), (Kuhn 1970b, p. 241), as well as Chapter 5 of this thesis.

> . . . suppose that we had a generally followed set of rules for rational pursuit. Then all rational scientists, or at least almost all, would make the same decision concerning the choice of a theory to work on. This would destroy an essential condition for progress within scientific community. . . . If each member of the community would follow the rule for pursuit, there would just be *one* preferred theory for the whole group to work on. Rational behavior of this sort could not produce the diversity of research that seems important to scientific progress. (Rueger 1996, p. 265)

There are two problematic assumptions present in Rueger's argument. First, he assumes that a set of rules for rational pursuit necessarily has to result in only one theory being worthy of pursuit. Second, he assumes that if a theory is assessed as worthy of pursuit, then the entire scientific community in the given domain is to actually pursue it. A possible reason why he makes these two assumptions is that he fails to distinguish between the rules for rational pursuit that hold for a community, and those that hold for an individual scientist. While an individual scientist is more likely to find one theory as worthy of pursuit for herself or himself, there is no reason why the same should follow for the scientific community in the given domain. Moreover, while a theory may be worthy of pursuit for the community, that does not imply that it is worthy of pursuit for each member of this community.

Finally, the neglect of this distinction has lead some authors to argue that non-epistemic factors are crucial for the diversity regarding the division of labor among scientists. For instance, Kitcher (1990) suggests that

> we sometimes want to maintain cognitive diversity even in instances where it would be reasonable for all to agree that one of two theories was inferior to its rival, and we may be grateful to the stubborn minority who continue to advocate problematic ideas. (p. 7)

As an example, Kitcher mentions Alfred Wegener's theory of continental drift, and asks:

> Was it equally reasonable to be a drifter or an anti-drifter in the 1920s and 1930s? Inspired by appreciation of the intricate shifts in standards of appraisal that occur in the history of science, you might say "yes". But then you face a problem of maintaining cognitive diversity . . . from the community's point of view, it would have been better if geologists had been more equally divided. (p. 7-8)

Furthermore, Kitcher maintains that the problem of the cognitive diversity cannot be solved by introducing the stance of pursuit:

> Once we have recognized the distinction [between belief in a theory and pursuit of research designed to apply or extend that theory], can we not accept a simple solution to my puzzle? Whereas it may

> be rational for each scientist to believe the theory that is better supported by available evidence, it may not be rational for each of them to pursue that theory, and what the community cares about is the distribution of pursuit not the distribution of belief. ... The idea that it is rational for a person to believe the better-supported theory seems, however, to be based on supposing that that person's aim is to achieve true beliefs (or some other desirable epistemic state, the acceptance of empirically adequate theories, for example). In that case, however, it appears that the person should also pursue the better-supported theory, since pursuing a doctrine that is likely to be false is likely to breed more falsehood (or less of the desired epistemic state). Only if we situate the individual in a society of other epistemic agents ... does it begin to appear rational for someone to assign herself to the working out of ideas that she (and her colleagues) view as epistemically inferior. (p. 8)

Kitcher's solution is to conceive of diversity of the pursuit in the given scientific community by means of personal motives and differences that scientists have in their assessments of epistemic merit.

The key problematic assumption endorsed here is that the assessment of pursuit worthiness for the given community will necessarily give only one "better-supported" theory. Hence, Kitcher oversees the fact that more than one theory can be simultaneously worthy of pursuit for the given community. More precisely, what he oversees is the fact that *an individual scientist* may assess a theory as worthy of pursuit *for the community*, and yet, engage in a pursuit of a different theory, which is also worthy of pursuit for the given community (and probably for the scientist as well). To go back to the example of Wegener's theory, the pursuit worthiness of the theory of continental drift for the geological community in the 1920s and 1930s could be perfectly justified by means of evaluation that holds for the community (that is, without introducing non-epistemic factors regarding individual scientists and their preferences), as we will show in Chapter 4.[11]

In view of these three examples[12] we can conclude that not only has a very simple thought – that more than one theory can be worthy of pursuit at the same time – been neglected, but the distinction between individual and group rationality underlying the context of pursuit has been insufficiently clarified.

[11] Kitcher's point that non-epistemic reasons govern the diversity may very well be descriptive of the actual scientific practice. However, our aim here was to point out that it does not have a normative strength, since diversity could be maintained by evaluating the communal pursuit worthiness as well.

[12] A similar point could also be made for Richard E. Grandy's argument for diversity of pursuit (see (Grandy 2000)), as well as for Kuhn (see Chapter 5).

2.5 The Notions of Pursuit and Pursuit Worthiness Employed in this Thesis

Having made the above distinctions between different types of pursuit and different types of pursuit worthiness, let us clarify the notions that will be our primary focus in the remainder of this thesis.

First, we will be primarily interested in the pursuit of scientific theories (rather than the pursuit of phenomena, entities, technological developments, etc.). Second, we are interested in investigating the notion of epistemic pursuit worthiness (rather than what we have called the practical pursuit worthiness). Finally, such a notion of epistemic pursuit worthiness will regard what is worthy of pursuit for the given scientific community (rather than what is worthy of pursuit for an individual scientist).

Chapter

3 Epistemic Justification in the Context of Pursuit

A Coherentist Approach

✎ *This chapter is based on a paper with the same title, which is accepted for publication in a special issue of Synthese "Is Science Inconsistent?", ed. by P. Vickers and O. Bueno (Šešelja and Straßer 201x). The paper is a result of a joint work with Christian Straßer. We are indebted to Erik Weber and the anonymous reviewers for valuable comments on a former draft of this paper.*

Summary The aim of this chapter is to offer an account of epistemic justification suitable for the context of theory pursuit, that is, for the context in which new scientific ideas, possibly incompatible with the already established theories, emerge and are pursued by scientists. We will frame our account paradigmatically on the basis of one of the influential systems of epistemic justification: Laurence Bonjour's coherence theory of justification. The idea underlying our approach is to develop a set of criteria which indicate that the pursued system is promising of contributing to the epistemic goal of robustness and of developing into a candidate for acceptance. In order to realize this we will (a) adjust the scope of Bonjour's standards –consistency, inferential density, and explanatory power–, and (b) complement them by the requirement of a programmatic character. In this way we allow for the evaluation of the "potential coherence" of the given epistemic system.

3.1 Introduction

In this chapter we present a coherentist account of epistemic justification suitable for the evaluation of the epistemic pursuit worthiness of scientific theories. Among the accounts of epistemic justification that have been applied to the evaluation of scientific theories, coherentist approaches have been scoring fairly

better than the foundationalist ones.[1] Thagard's explanatory coherentism has been applied to a number of scientific revolutions (Thagard 1992), Chang has applied his epistemic iteration to the case of the invention of the concept of temperature (Chang 2004), while Bonjour's coherentism (Bonjour 1985, 1989) has been discussed with regard to the problem of theory choice in Kuhn's philosophy of science (Kuukkanen 2007). Nevertheless, none of these accounts is fully suitable for theory evaluation in the context of pursuit. Thagard's explanatory coherence is primarily concerned with theory acceptance, even when it is applied to the cases of early developments of theories (e.g. see his discussion of Wegener's theory of the continental drift (Thagard 1992, p. 171)). In contrast, in Chang's epistemic iteration coherence is used "as a guide for a dynamic process of concept formation and knowledge building, rather than strict justification" (Chang 2004, p. 156). Even though his approach could be seen as addressing epistemic justification in a broader sense of the term, it does not offer any criteria for an assessment of the initial pursuit worthiness of theories, which, as we shall see, is an important part of the evaluation of pursuit worthiness. Bonjour's coherentism offers criteria of epistemic justification as it has been traditionally conceived, that is, regarding the context of acceptance (we will discuss this more in Section 3.2).

As we have mentioned in Chapter 1, a theory is epistemically worthy of pursuit to the extent that it can be shown to have a promising potential for contributing to the epistemic goals of the scientific enterprise. In a coherentist framework, we can say that a theory is epistemically worthy of pursuit to the extent that it can be shown to have a promising potential for contributing to a greater coherence and greater robustness of scientific knowledge. Such a character can be manifested in the theory being promising of increasing the coherence of an already existing research tradition (for example, by deepening its explanatory mechanisms), and/or in the theory being promising of developing into a system or a research tradition that is more coherent than the currently established one in the given domain, and/or in the theory being promising of developing into a good backup theory or a research tradition for the currently dominant one.[2] The latter feature is motivated by the epistemic virtue of robustness of knowledge.

The significance of robustness as an epistemic goal becomes even more obvious once we have stepped on the coherentist ground. In order to explain why, let us briefly recall that the main virtue of a foundationalist approach to epistemic justification is the "firm foundation" of knowledge: it is supposed to be not only the basis of epistemic justification, but also to offer robustness to the knowledge, since even if our theories turn out to be wrong, there is

[1] For example, see the discussion in (Kleiner 2003, p. 513-514) and (Chang 2004, p. 223-224).

[2] In the remainder of the chapter we will – for the sake of simplicity – use the terms "cognitive system", "(scientific) theory" or "scientific hypothesis" interchangeably. It is clear though that especially in the early stages of their development, such cognitive structures have neither all the properties of a theory nor all the links which would make them sufficiently systematic, and yet, they can be more than just a hypothesis.

always a firm ground to get back to and start building the knowledge all over again. To use Neurath's analogy mentioned in Chapter 1, a foundationalist approach presupposes that there is always a dry-dock where the ship can be reconstructed on the basis of its firm components all over again. Needless to say, the idea of such an unquestionable basis of scientific knowledge has by many been regarded untenable, and coherentism stepped in as an alternative. Now, in order to compensate for the lack of a firm foundation, coherentism, applied to scientific knowledge, needs to introduce robustness in a different way, namely, by allowing for developing theories to serve as backups of the currently established ones. In other words, if there is no absolutely certain foundation to which we can always turn to, then the best we can do in order to achieve robustness of the scientific knowledge as a whole with respect to the uncertainty of the future developments, is to allow for a pursuit of theories that are alternative to (and possibly incompatible with) the currently established ones.

In summary, our task is to offer the standards on the basis of which it can be judged to which degree a theory is epistemically worthy of pursuit, viz. to which extent it is potentially epistemically justified viz. *potentially coherent*. We will paradigmatically build our framework on the basis of Bonjour's account. The advantage of adjusting an already existing theory of justification is that we can obtain a single unifying (in this case, coherentist) framework of justification, covering both the context of pursuit and the context of acceptance. The reason why we have chosen Bonjour's criteria for this purpose is that, on the one hand, they are concise and simple, which makes them especially suitable for a demonstration of our approach. On the other hand, they are sufficiently similar to other approaches to theory evaluation (such as, for example, Laudan's problem-solving approach (Laudan 1977) or the approach of Kitcher's explanatory unification (Kitcher 1989) or Thagard's explanatory coherentism (Thagard 1992)). This means that our account can be easily adjusted to fit different methodological frameworks. The way we will modify Bonjour's criteria is by adjusting them in such a way that their focus is shifted towards those aspects of theories which point to their epistemic pursuit worthiness.

The chapter is structured as follows. In Section 3.2 we give a brief summery of Bonjour's system of coherence. In order to see which requirements should be satisfied for an epistemic evaluation to be adequate for the context of pursuit, we will in Section 3.3 present the main ideas of Laudan's notion of the context of pursuit. In view of Laudan's ideas, we will in Sections 3.4 to 3.6 present the criteria of potential coherence, applicable to the evaluation of scientific theories in the context of pursuit. In Section 3.7 we will give a meta-justification of our framework. Finally, Section 3.8 brings some concluding remarks.

3.2 Bonjour's Concept of Coherence

Bonjour defines coherence by means of the following criteria (Bonjour 1985, p. 95-99):[3]

1. *Consistency:*
   a) *Logical consistency:* A system of beliefs is coherent only if it is logically consistent.[4]
   b) *Probabilistic consistency:* A system of beliefs is coherent in proportion to its degree of probabilistic consistency. Probabilistic inconsistency occurs when a system of beliefs contains both the belief that P and also the belief that it is extremely improbable that P. According to Bonjour, probabilistic inconsistency differs from the logical one in two respects: (i) it is extremely doubtful that probabilistic inconsistency can be entirely avoided; (ii) probabilistic consistency (unlike the logical one) is a matter of degree, depending on how many such conflicts the system contains, and the degree of improbability involved in each case.
2. *Inferential density:*
   a) *Presence of inferential connections:* The coherence of a system of beliefs is increased by the presence of inferential connections between its component beliefs and in proportion to the number and strength of such connections.[5]
   b) *A lack of inferential connections:* The coherence of the system of beliefs is diminished to the extent to which it is divided into subsystems of beliefs which are relatively unconnected to each other by inferential connections.
3. *Explanatory power:* The coherence of a system of beliefs is decreased in proportion to the presence of unexplained anomalies in the believed content of the system. Bonjour defines an anomaly as a fact or event, especially one involving some sort of recurring pattern, which is claimed to obtain by one or more of the beliefs in the system of beliefs, but which is incapable of being explained (or would have been incapable of being predicted) by appeal to the other beliefs in the system.

---

[3] For the sake of transparency we will give each criterion (or group of criteria) an appropriate name.

[4] Bonjour remarks that making the criterion for consistency absolutely necessary might be an oversimplification. Moreover, recent research has shown that it is sensible to ask how inconsistent a theory is and that logical inconsistency can be considered to come in degrees as well. In order to measure such degrees syntactic approaches based on minimal inconsistent sets (Hunter and Konieczny 2008) or maximal consistent sets (Knight 2002) have been suggested, as well as semantic approaches employing paraconsistent models such as (Hunter 2002, Hunter and Konieczny 2005, Grant 1978, Grant and Hunter 2006, 2008, Ma et al. 2009).

[5] It is interesting to notice that William Wimsatt emphasizes the role of two more refined notions in scientific theories that are based on the inferential density, namely the robustness and the generative entrenchment of parts of cognitive systems. Given a directed graph of inferential connections "a robust node has multiple inferential paths leading *to* it and resists failure because of its multiple sources of support." and the generative entrenchment of a node is given proportional to the "number of nodes reachable from that node" (Wimsatt 2007, p. 142).

Conceived in this way, Bonjour's concept of coherence – to which we will refer to as *the actual coherence* – can be used for an assessment of different belief systems, including scientific theories.[6] Nevertheless, in order to be applicable to the evaluation of scientific theories, a few adjustments should be made to Bonjour's criteria:

1. First, Bonjour's unit of appraisal – "belief system" – should be replaced with "cognitive system". The reason for this is that, especially in the context of pursuit, a scientist does not necessarily have to believe in the assumptions constituting her model, but rather take them as provisional descriptions of what is yet to be further investigated. For instance, she may not believe that some concepts used in a certain pursued scientific model refer in a strict sense or that some ad hoc hypotheses are true.
2. Once we have adjusted the unit of appraisal, it becomes clear that the criterion of inferential density should be adjusted as well. In contrast to the inferential density of the belief system of an individual, which also addresses inferential connections between two theories belonging to it, the inferential density of a scientific theory can be seen as consisting of two aspects: the internal one – referring to the inferential connection within the theory itself, and the external one – referring to the connections between the theory and other scientific theories. The standard of inferential density, adapted for the evaluation of scientific theories, is now formulated in the following way:

   > The coherence of the cognitive system is increased by the presence of both, the inferential connections within the system, as well as the connections between the evaluated system and other established scientific systems; and vice versa, the coherence is diminished by the absence of the connections within the system as well as connections with other theories that are considered relevant for it (for example, if they have an overlapping explanatory scope, or if one theory is expected to deepen the other, etc.).

3. For the same reason, we can distinguish between the internal and the external aspects of the consistency criterion. Bonjour's formulation of the standard of consistency can be taken to refer to the internal consistency,[7] while the standard of the external (logical and probabilistic) consistency is now expressed in the following way:

   > A system of beliefs is coherent in proportion to its logical and probabilistic consistency with other, already established scientific theories.

Even though rooted in a coherentist framework, Bonjour's criteria are similar to some other standard approaches to theory evaluation. For example,

[6] "By devising a new system of theoretical concepts the theoretician makes an explanation available and thus enhances the coherence of the system. In this way the progress of theoretical science may be plausibly viewed as a result of the search for greater coherence." (Bonjour 1985, p. 100).

[7] For the precise distinction between the internal and external consistency, see Section 3.5.3.

Laudan's view on the scientific development as a progress in problem-solving employs similar criteria: explanatory power is expressed in terms of empirical problem-solutions, inferential density in terms of compatibility between theories (the absence of which is a type of external conceptual problems) and a preference for more unifying problem-solutions, while inconsistency falls under internal and external conceptual problems (Laudan 1977). In a similar vain, it could be argued for the similarity between Kuhn's and Bonjour's criteria. For example, as already mentioned, (Kuukkanen 2007) suggests that Kuhn's philosophical standpoint could be incorporated into a coherentist epistemology. Furthermore, according to Kitcher, "scientists in the thick of a controversy face two types of predicaments: those of inconsistency and those of explanation." (Kitcher 2000, p. 31). If we add to that Kitcher's view on the aim of inquiry as "the provision of a maximally unified set of explanatory schemata that will generate the largest possible set of true instantiations" (Ibid., p. 24), we can see a clear similarity between Bonjour's and Kitcher's evaluation criteria. Therefore, rooting our account in a coherentist epistemology and moreover, in Bonjour's notion of coherence, should not prevent it from being relevant for other approaches to theory assessment as well. It should rather be seen as a paradigmatic demonstration of adjusting existing accounts of epistemic justification to the context of pursuit.

3.3 Laudan on the Context of Pursuit

Laudan distinguishes two contexts in which cognitive appraisals of scientific theories and research traditions are made:

1. *The context of acceptance* is a context in which scientists choose to accept one of the competing theories (and research traditions), thus treating it as if it were true (Laudan 1977, p. 108).
2. *The context of pursuit* is a context specific for the emergence of a new research tradition, in which scientists begin to pursue and explore it long before it is qualified to be accepted over its older rivals (ibid., p. 110).

By recognizing the context of pursuit, we can explain the historical fact that scientists sometimes, particularly in the time of scientific revolutions, work alternatively in two different, even mutually inconsistent, research traditions:

> I shall suggest that, within each of these contexts of inquiry, very different sorts of questions are raised about the cognitive credentials of a theory, and that much scientific activity which appears irrational – if we insist on a uni-contextual analysis – can be perceived as highly rational if we allow for the divergent goals of the following two contexts: [the context of acceptance and the context of pursuit] (ibid., p. 108).

Thus, accepting one research tradition does not necessarily have to preclude scientists from pursuing alternatives which are inconsistent with it. For exam-

ple, Galileo's physics was in its early stage much weaker than its primary rival, Aristotelianism:

> Aristotle's research tradition could solve a great many more important empirical problems than Galileo's. Equally, for all the conceptual difficulties of Aristotelianism, it really posed fewer crucial conceptual problems than Galileo's early brand of physical Copernicanism ... Galileo was taken seriously by later scientists of the seventeenth century, not because his system as a whole could explain more than its medieval and renaissance predecessors (for it probably could *not*), but rather because it showed promise by being able, in a short span of time, to offer solutions to problems which constituted anomalies for the other research traditions in the field. (ibid., p. 112, italics in the original).

A similar case is Daltonian atomism. Dalton's early atomic theory was confronted by numerous serious anomalies, and was far from reaching the problem-solving success of its dominant rival – elective affinity chemistry. Still, his system was taken to be promising since, in contrast to its rival, it was able to predict what has later on come to be known as the laws of definite and multiple proportions (namely, that chemical substances combine in certain definite ratios and multiples thereof no matter how much of the various reagents was present (ibid., p. 113)). Dalton's theory was thus acknowledged as worthy of further pursuit mainly because of its scientific promise, despite the fact that it did not (yet) satisfy the standard of theory acceptance.

Taking these examples into account, it is obvious that Bonjour's criteria of coherence aren't suitable for the context of pursuit. Both Galileo's and Dalton's theories were in the early stages of their development less coherent than their dominant rivals. Still, both of them were taken to be worthy of pursuit. An account of epistemic justification applicable to the context of pursuit should be able to offer the standards in view of which we can understand why their pursuit was rational.

Before we present our account of coherence evaluation for the context of pursuit, let us make a critical remark on Laudan's approach. Laudan gives the following criterion for when it is rational to pursue a research tradition: "*it is always rational to pursue any research tradition which has a higher rate of progress than its rivals* (even if the former has a lower problem-solving effectiveness)" (ibid., p. 111, italics in the original). Since only one tradition at the time can have a higher rate of progress than its rivals, Laudan's criterion can evaluate pursuit worthiness of only one tradition in cases in which different rivaling traditions are simultaneously worthy of pursuit. Indeed, if a tradition is worthy of pursuit, that does not imply rejecting its rivals as unworthy of pursuit. There are situations in which it may be rational for a given scientific community to pursue two or more research traditions at the same time. One of these traditions may exhibit a higher rate of progress than its rivals at one point, but soon it may turn out to be the other way around, or they may

exhibit similar rates of progress.[8]

Our idea of pursuit worthiness is in this respect more similar to the one developed by Laurie Anne Whitt (1990, 1992), whose indices of theory promise also allow for more than one theory at the same time to be evaluated as being promising of further investigation (Whitt 1992, p. 632).[9]

3.4 The Notion of Potential Coherence

In order to clarify the idea underlying our account of potential coherence, let us make an analogy between a scientist pursuing a theory and a painter trying to paint a particular landscape in a more realistic way than other painters have done it so far. Just like the painter will start with a simple sketch, so does the scientist begin with an abstract model. And just like the sketch is far from being the final form of the painting, ready to compete with other already finished paintings, so is the pursued theory in its beginnings not able to compete with its dominant rival with respect to its actual coherence. Nevertheless, the sketch could already show certain strengths due to which we can say that it seems promising of *becoming* as realistic a painting of the landscape as other works, if not even more. For example, even though it is still lacking various nuances of colors, a number of details, etc., we could imagine that the previous painters did not have a technique for representing three-dimensional objects, whereas the new painter develops such a technique and thus introduces a significant novelty allowing for a more realistic depicting of the landscape than it was possible earlier. And so, when evaluating the promising character of this work, rather than criticizing it for being a sketch with many shortcomings compared to the finished paintings, we would focus on its strengths and particularly on those elements which the other paintings have not managed to include. As the painting develops we can evaluate whether it is still promising or whether those initially interesting elements have ceased to be interesting (for example, another painter could have in the meantime incorporated the novelty into an already more realistic painting, while our painter did not manage to improve any other elements of her work.) In a similar way, the criteria of the potential coherence should enable us to judge whether a developing scientific theory is sufficiently promising in spite of its obvious shortcomings.

When we say *potentially* coherent, we are not after a degree of coherence which a theory will certainly have in the future. It is clear that we cannot

[8] Alexander Rueger's criticism of Laudan's criterion (Rueger 1996, p. 267) along the similar lines overlooks the fact that Laudan expresses his criterion only as a sufficient, but not a necessary one. In contrast, according to (Whitt 1992, p. 616-617), Laudan's criterion is not even sufficient for the evaluation of pursuit worthiness.

[9] Note that even though the criteria constituting our account are similar to those explicated by Whitt, our approaches differ in several key respects. First, our account is formulated in terms of a coherentist account of epistemic justification, while Whitt's approach is rooted in Laudan's (Laudan 1977) and McMullin's (McMullin 1976) methodological frameworks. Second, our account introduces a unificatory aspect to the evaluation in the context of pursuit and the context of acceptance by allowing for both to be presented within the same epistemic framework (namely, Bonjour's coherentism).

foresee that with certainty. What we are after are indications which tell us that the theory *might* develop or *is promising* of developing into a powerful scientific system. The fact that we speak here of a possibility, instead of a full guarantee, means that the result of such an evaluation will always include a certain level of chance and risk. We are thus interested in certain properties, a theory can *actually* hold, on the basis of which experts in the given field are able to say whether taking such a risk is epistemically justified or not. In order to account for this we take into account on the following two aspects:

1. In order to evaluate the potential coherence of T we need to focus on certain positive features of the theory, which highlight that there are good reasons to consider it to be an (epistemically) attractive candidate for further pursuit. These are often properties which the dominant rival lacks. The idea is here to restrict the focus of the criteria of the actual coherence (in our case, Bonjour's original criteria, see Section 3.2) to the particular strengths of the theory, which should serve as indications of whether the theory is sufficiently promising. We will thus assess the *potential* explanatory power of the theory, its *potential* inferential density and its *potential* consistency.
2. Since we cannot expect a theory to be fully developed right from the start we should not be too critical of the various shortcomings that it faces. However, we should also not turn a blind eye on them since we want to evaluate if the theory in question is promising of overcoming these problems. What we expect instead is that T offers a *programmatic character*, i.e. methodological and heuristic means to tackle these problems in its further development.

The diachronic character of pursuit points to two distinct questions concerning the assessment of pursuit worthiness. First, we may ask whether a theory is *initially* worthy of pursuit. The importance of this question can easily be seen in the case of assessing new research proposals with which scientists apply for funding. Even though certain non-epistemic factors also play a role in such an assessment, proposals for the research of new scientific theories should, in principle, be attractive from an epistemic perspective as well. In other words, we are usually interested in financing the research of theories that are epistemically worthy of pursuit. However, even if a theory is initially promising, that does not guarantee that it will also remain worthy of further pursuit. Therefore, we will show how our criteria of the potential coherence are applied when evaluating theories in these two respects.

The *initial* pursuit worthiness will be assessed by means of the following criteria:

C1. Potential Explanatory Power

C2. Potential Inferential Density

C3. Potential Consistency

C4. Programmatic character.

For an assessment of the *successive* pursuit worthiness, we will use the same set of criteria, but in addition, the criteria will require a gradual strengthening, which takes into account our expectations concerning the growth of the theory.

It is important to notice that even though it is difficult to pinpoint the exact moment when a theory should be subjected to the evaluation of its initial pursuit worthiness, principally speaking, we are referring to the moment when it enters the discourse of the given scientific community (for example, by being presented in a publication or at a scientific conference). This may also be a time period during which the idea has been proposed and has received the initial criticism. Moreover, the distinction between the initial and the subsequent pursuit worthiness has to be done in view of the specific historical and scientific context. For example, the subsequent pursuit worthiness could refer to the subsequent models of the given theory, or the subsequent theories in the given research tradition.

In the following two sections we will introduce each of the above mentioned criteria. We will then in Section 3.7 give a meta-justification for them. While the aim of Bonjour's meta-justification is to show that his system is truth-conducive, the task of our meta-justification is to show that our criteria are conducive – as much as that is possible in the context of pursuit – of the actual coherence and the epistemic goal of robustness.

3.5 Initial Pursuit Worthiness

The initial pursuit worthiness is assessed in terms of the criteria C1–C4. Since we are interested in possible future developments of a theory and possible ways in which it can contribute to the development and robustness of our scientific knowledge as a whole, the evaluation of the pursuit worthiness of a theory needs to be done in view of already established scientific findings. In other words, the theory is to be evaluated against the *cognitive horizon* characterized inter alia by the more entrenched theories of its time, which it may in part inherit and in part challenge. The cognitive horizon includes, for instance, questions such as what is considered as an explanandum in the given domain, what are important problems, what is a proper methodology for the research in the given domain, etc.

3.5.1 Potential Explanatory Power

We begin by discussing first Bonjour's criterion regarding explanatory connections in the evaluated system of beliefs. According to Bonjour's requirement, the coherence of the system is decreased in proportion to the presence of unexplained anomalies. However, as we have seen in Section 3.3 (for example, in the case of Galileo's physics and Dalton's chemistry), what we need to focus on in the context of pursuit is what the system can actually explain or predict[10]

[10]Since Bonjour treats explanations and predictions as involving the same sort of inferential relations (see (Bonjour 1985, p. 240, Note 15)), we will do the same. It would be possible,

and on the question of how significant those explanations and predictions are, in spite of there being a number of unexplained anomalies. Therefore, we need to introduce a weaker version of the requirement for explanatory power:

> *Potential Explanatory Power*: The potential coherence of a cognitive system is increased by the presence of explanations that are considered to be significant at that point of the scientific development. An explanation[11] is significant if:
>
> a) it addresses the phenomena for which the established or other pursued rivals have either no explanation, or have explanations which are weak (a weak explanation would be, for example, an explanation that introduces new conceptual problems);
> b) it addresses certain benchmark problems or questions in the given scientific domain in a novel way.

For example the fact that classical thermodynamics could not solve the problem of the blackbody radiation, made explaining this phenomenon significant. Or, for instance, during the development of Galileo's physics a number of anomalies appeared for the Aristotelian framework. Even though the proponents of the latter one offered explanations of these phenomena, they "smacked of the artificial and the contrived" (Laudan 1977, p. 112). Finally, Copernicus' heliocentric system provided an explanation (or rather a prediction) of planetary movements from the assumption of a non-stationary Earth. In this case, the explained phenomena were not observations that were unexplained in the Ptolemaic system.[12] Rather, Copernicus offered a new way of explaining some of the most important phenomena in the sixteenth century astronomy.

Let us conclude this section by discussing the epistemic status of explanations of young theories. Explanations of young pursued theories have often a certain *prima facie* or hypothetical character which may be, for instance, due to their idealized nature, due to imprecision of the measurement of data supporting them, or due to the fact that the epistemic status of (some of) the evidence supporting the explanations is itself in need of investigation. In contrast, although we probably never reach a state in scientific development in which explanations are final and "all-things-considered", the explanations of an accepted theory are "enough-things-considered" so that we characterize them as epistemically justfied, together with the theory to which they belong.[13]

---

of course, to introduce a separate criterion for the predictive power of a theory.

[11] Even though we are not here discussing the notion of a scientific explanation, it is the task of an account of explanation fitting our model to be able to dismiss spurious explanations as non-scientific (e.g., if someone offers to "explain" all the phenomena by claiming that they occur because god wanted them that way). For instance, in view of a causal-mechanical account, most of spurious explanations can be rejected due to the fact that they do not offer any underlying causal mechanism.

[12] As Friedel Weinert remarks: "Copenicus' observations do not establish any *new* facts. ... It is therefore fair to say that from an observational point of view, the Copenican and Ptolemaic systems were equivalent." (Weinert 2009, p. 24-25, italics in original).

[13] Of course, "How much is 'enough'?" is one of the essential questions of epistemic justification in the context of acceptance.

For instance, Miller (2002) argues that a certain ignorance concerning observational inaccuracies in experiments or experimental data is sometimes fruitful for the explanatory strength and growth of young scientific theories. As an example he brings Galileo's thought experiment according to which, when a stone is dropped from a moving ship, it will drop directly under the person who has dropped it. Such a conclusion was in disagreement with Aristotelian assumptions. But had Galileo "extremely precise measuring instruments" (Miller 2002, p. 36), then an empirical experiment would not have had the predicted outcome. Various forces (such as the Coriolis force, the centrifugal force, etc.) would have interfered Galileo's idealized setting based on an inertial reference system and a free fall through vacuum. Similarly was Einstein's special theory of relativity based on the idealized assumption "of an inertial reference system in the sense that it only deals with measurements made in such systems, which Einstein himself took to be its 'logical weakness'" (ibid., p. 37).

Another example is the case of Galileo's telescope, which allowed for important discoveries and an increase in the explanatory power of the heliocentric view. Initially there was no theory of lenses available to epistemically justify the evidence gained by means of telescopes. Only later, around 1610, did "Kepler [...] have a knowledge of optics [...] which enabled him to [...]the *theory* of the telescope." (van Helden 1974, p. 40).[14] However, as van Helden points out, "the importance of the new discoveries [by means of telescopic observations] was tremendous", they " were bombshells indeed" and although "the opposition was powerful, and the instruments were very poor", "the time it took Galileo to convince all reasonable men was astonishingly short" (ibid., p. 51). One reason was that Galileo was able to significantly improve the telescopes (see (ibid., p. 52)). Moreover, telescopic observations were objective in the sense that they were reproducible. Additionally "the telescope was instrumental in the growth of the idea that the laws of nature apply everywhere equally [...] everywhere in the universe" (ibid., p. 57) and thus it contributed to a more unifying scientific worldview.

The fact that idealized settings, though being "less accurate" in terms of data, are "more informative" often serves as a catalyst for a further fruitful development in which the empirical precision step by step gets improved by more and more accurate models. The same goes for ad hoc hypotheses, arguments by analogy, the (temporary) acceptance of inconsistencies (see Section 3.5.3) and similar epistemic tools that are often useful to master early obstacles of theory building. Only later in Einstein's general theory of relativity the approximative character of his earlier model has been overcome by allowing for accelerating reference systems. Similarly, Galileo's discoveries served as a catalyst for the study of lenses which "became an important part of optics" (ibid., p. 52), and which in the long run deepened his cognitive system by epistemically justifying the evidence gained by means of telescopes.

[14] We are indebted to Steffen Ducheyne for suggesting van Helden's paper to us.

3.5.2 Potential Inferential Density

The potential inferential density is evaluated in terms of the potential internal inferential density and the potential external inferential density.

3.5.2.1 Potential Internal Inferential Density

It is easy to notice that the more phenomena can be explained by the same set of hypotheses, the more unifying the system gets. For example, the underlying idea of Kitcher's account of explanatory unification is that the same argument pattern is to be used in the derivation of a wide range of explananda (Kitcher 1981a, p. 512). In a similar way Thagard suggests that "we should prefer theories that generate explanations using a unified core of hypotheses." (Thagard 1992, p. 67). Now, to ask how inferentially dense a system is, is in fact to ask, how unified it is (Bonjour 1985, p. 97). Therefore, if a theory which is being evaluated for its potential coherence, has a more unified core of hypotheses than its rivals, this will contribute to its potential inferential density. More precisely:

> *Potential internal inferential density:* The potential coherence of the cognitive system is increased by the presence of a unified core of hypotheses.

For example, during the revolution in the geological sciences, Hess' theory of seafloor spreading which explained the mechanism of the continental drift – offered a highly unified explanation of the phenomena relevant to the ocean basins. Hess' main hypothesis could account for a wide range of phenomena, in contrast to permanentist and contractionist theories that required additional ad-hoc hypotheses to explain the same set of explananda (Le Grand 1988, p. 198).

There are different accounts of explanatory unification (e.g. (Friedman 1974, Kitcher 1981a, Weber 1999)) that could be applied for a more precise assessment of unifying properties of scientific theories.

3.5.2.2 Potential External Inferential Density

If the theory in question shows certain inferential connections with other established or pursued theories, for example, with those from different scientific domains, this will speak in favor of its potential coherence. An example of such inferential connections would be analogical relations between theories: one theory can develop using an analogy between its explananda and the explananda of another, already developed theory. According to Laudan, analogy is, in fact, the most common form of mutual reinforcement between theories (Laudan 1977, p. 230). For example, Huygens used the analogy between the familiar phenomena of water and sound waves and the hypothetical phenomena of the light waves to explain the nature of the latter ones (Thagard 1981, p. 249). Moreover, if the theory establishes inferential connections with theories with

which the dominant rival has not managed to connect so far, this will increase its potential inferential density. In this way a young theory shows the promise of contributing to the aim of robustness: it points out some of the advantages it can introduce in case the dominant rival turns out to be a weak theory after all. We can thus formulate the following criterion:

> *Potential external inferential density:* The potential coherence of a cognitive system is increased by the presence of significant inferential connections between this system and other established or pursued scientific theories. The connection is significant if:
>
> a) it links the system to a theory with which a dominant or a pursed rival has not managed to connect so far;
> b) it links the system to a theory that is itself considered established.

For example, the fact that Galileo's physics had inferential connections with Kepler's astronomy and Copernicus' heliocentric system, contributed to the promising character of all three theories.

3.5.3 Potential Consistency

Finally, the criterion of consistency needs to be modified. The following discussion will refer to both logical and probabilistic inconsistencies. First of all, if we make a distinction between a theoretical framework and observations, we can make a difference between, on the one hand, inconsistencies within a theory or between theories, and on the other hand, inconsistencies between a theory and observations (see (Priest 2002, p. 122)).

We first discuss those within or between theories. Taking into account the place in the theoretical framework where the inconsistency appears, we can distinguish between:

1. *Internal inconsistency*, which concerns inconsistencies within the theory. We say that a theory[15] is *logically* internally inconsistent if it contains a proposition and its negation. Using Bonjour's notion of probabilistic inconsistency (see Section 3.2), we can say that a theory is *probabilistically*

[15] We take a theory to consist of a certain set of propositions and all their consequences. We restrict the latter to the consequences which are known at the given time point since the coherence evaluation can only take into account the known consequences and thus the inconsistencies known at that time point. On the basis of new consequences it may turn out that a seemingly consistent theory is in fact inconsistent. Two other remarks are important at this point. First, we are of course aware of the fact that the language of propositional logic is rather suboptimal for the task of a proper formalization of scientific theories. We use it as a simplification (and so does Bonjour). It is an open question what formal language is best for this task, and if there is any one (!) such language at all. Second, it is often the case that especially (but not only) immature theories contain contradictions. To speak about consequences of such systems in terms of classical logic is not very helpful, since in face of contradictions classical logic derives anything. Logicians have developed various ways to cope with such situations by use of paraconsistent logics (see e.g. (Béziau and Carnielli 2006, Batens et al. 2000)).

internally inconsistent if it states P and also that it is highly improbable that P.

2. *External inconsistency*, which concerns inconsistencies (logical or probabilistic) between the evaluated theory and other, already established theories. We say that a theory is *logically* externally inconsistent if it contains a proposition, while some other, established or pursued theory contains the negation of this proposition. A theory is *probabilistically* externally inconsistent if it states P while some other, established theory states that it is highly improbable that P (or the other way around). We can further make a difference between two types of external inconsistencies:
   a) a theory T_1 is inconsistent with an established rival theory T_2;
   b) a theory T_1 is inconsistent with an established theory T_2 belonging to a different scientific domain.

3.5.3.1 Potential Internal Consistency

It is important to notice that it is an oversimplification of the scientific practice to treat the criterion for consistency as absolutely necessary (see (Bonjour 1985, p. 247)). The higher the potential explanatory power of a theory is, the more forgiving we are towards possible inconsistencies in it. More precisely, by temporarily accepting certain inconsistencies, we may enhance the overall potential explanatory power of the system. As Thagard points out: "It may turn out at a particular time that coherence is maximized by accepting a set A that is inconsistent, but other coherence-based inferences need not be unduly influenced by the inconsistency, whose effects may be relatively isolated in the network of elements." (Thagard 2000, p. 74-75).[16] In a similar vain, Lakatos suggests that "it may be rational to put the inconsistency into some temporary *ad hoc* quarantine, and carry on with positive heuristic of the programme. This has been done even in mathematics, as the examples of the early infinitesimal calculus and of naive set theory show" (Lakatos 1978, p. 58).

Another example is Bohr's theory of the atom, which included both classical electrodynamic principles as well as quantum principles, which were mutually inconsistent. Bohr's idea was to temporarily ignore this inconsistency and to proceed on the inconsistent foundations (Lakatos 1978, p. 55). However, the developing theory was not abandoned due to this: "Though the theory was certainly considered as problematic, its empirical predictions were so much better than those of any other theory at the time that it had no real competitor." (Priest 2002, p. 123).

Thus, we should not consider it an initial drawback if a young research is still burdened by internal inconsistencies. However, since we strive for a high

[16] Even though Thagard doesn't clarify what he means by an inconsistency being *isolated* in the network of elements, he probably refers to avoiding the explosion which is inevitable in terms of classical logic. Some logics that are able to deal with inconsistencies block problematic applications of certain rules, such as disjunctive syllogism, to inconsistent formulas while at the same time allowing for the full classical derivative power for the consistent parts of a theory (see (Batens and Meheus 2006, Batens 1999)). In this sense they isolate inconsistencies since they prevent them from spreading.

actual coherence in the long run, it is a positive property of a cognitive system if it provides means to tackle these problems in its further development (see Sections 3.5.4 and 3.6). Of course, in the fortunate case that a young theory is internally consistent, this contributes to its attractiveness as does any problem that is avoided in first place and hence does not have to be tackled in the further development. Thus, we can formulate the following criterion:

> *Potential internal consistency:* The potential coherence of the cognitive system is increased if the system is consistent.

3.5.3.2 Potential External Consistency

The external inconsistencies of the first type – between a new and a rival, already established theory – have often been a reason for suspicion towards the former one. Just take the example of Copernicus or Galileo whose theories had some major inconsistencies with assumptions of the Aristotelian/Ptolemaic physics, which caused a serious resistance of the scientific community towards the acceptance of the new framework. Rejecting a pursued theory on the basis of such inconsistencies would make our assessment too conservative in character. Moreover, if the evaluated theory has a high potential explanatory power, it might very well turn out that it could one day become more (actually) coherent than its rival. In this case, the assumptions of the latter one would cease to belong to the accepted scientific views, and consequently, these inconsistencies would become irrelevant. That is why inconsistencies of this type do not necessarily have to be resolved by reconciling the opposing theories, but they could instead become obsolete, due to the winning of one of the rivals. Therefore, they should not be burdening for a theory which is assessed in the context of pursuit.

This also means that there is no reason why we should favor a theory that is consistent with the currently established one in the same domain (representing, for example, a possible development in the same research tradition). Allowing for both - developments that are consistent with the currently established theories, and those that are not, and that may even introduce new research traditions, are equally important for achieving the epistemic goals of coherence and robustness.

Let us now take a look at the second type of external inconsistencies – those between two theories belonging to two different domains. First of all, the fact that a pursued theory is inconsistent with some other scientific systems is not a sufficient reason for saying that it has a low potential coherence. For example, even though quantum mechanics and general relativity – both belonging to different subdomains of physics – are mutually inconsistent, neither has been rejected due to this inconsistency.

However, if a pursued theory turns out to be consistent with an established or pursued theory from another domain with which the dominant rival is inconsistent, this will speak in favor of its promising character. For example, if quantum mechanics had a rival theory T which were consistent with general

relativity, the potential external consistency of T would be increased. Similarly to the case of the external inferential density, consistency with such theories indicates that the pursued theory is promising of contributing to the robustness of scientific knowledge in the given domain. We can thus formulate the following criterion:

> *Potential external consistency:* The potential coherence of a cognitive system is increased if the system is consistent with significant theories. A theory is significant if:
>
> a) it is a pursued theory with which the rival of the evaluated theory is inconsistent, or
> b) it is an established theory in a different domain.

Just like in the case of internal inconsistencies, the (temporary) acceptance of external inconsistencies may be fruitful. For example, Adam Smith's economic theory was considered to be incompatible with the Newtonian thesis of a balance of forces in nature, since on the one hand, it relied on a general Newtonian balance of nature, while on the other hand, it postulated forces of economic motivation (e.g. self-interest) which were seemingly incompatible with such a balanced system (Laudan 1977, p. 230). Nevertheless, this inconsistency did not lead to the rejection of Smith's account, since the explanatory power of the theory was sufficiently high. In fact, the inconsistency became obsolete due to a conceptual change in view of which the notion of a force in economics became distinguished from the one used in physics. As a contrasting example, take Velikovsky's explanation of some historical events (such as the reported parting of the Red Sea for Moses), which was based on the hypothesis that Venus passed near the Earth some 5000 years ago. Velikovsky's theory was rejected not merely due to its inconsistency with Newton's law of motion (Thagard 1992, p. 91), but because it did not exhibit a high potential explanatory power. The phenomena it attempted to explain (namely, mythological events) could not even be considered as historical explananda in the proper sense. Furthermore, it did not have a sound and systematic methodology (see e.g. (Fitton 1974)), which would allow for an adequate programmatic character. Therefore Velikovsky's theory did not have the properties that could compensate for its external inconsistencies, and which would render it worthy of pursuit.

3.5.3.3 Consistency with Observations

Finally, let us have a look at (in)consistency of a theory with respect to observations. Observational inconsistencies may be viewed as being addressed by the explanatory properties of a theory. For instance, according to Graham Priest, inconsistencies of this type are viewed as explanatory anomalies (Priest 2002, p. 122), which is also the case according to Bonjour's definition of anomalies (see Section 3.2). When we talk about observations with which the theory is consistent then we restrict the focus on relevant observations: of course, in the majority of cases it is not of epistemic interest whether a scientific theory is

consistent with contingent facts such as the fact that we are right now working on this paper. Of epistemic interest here is, for instance, the evidential support for the theory in question, such as successful predictions that fall within this category. We have already mentioned that Bonjour treats these as being part of the explanatory scope of a theory (see footnote 10). Moreover, it can be argued that there is a positive feedback loop between an observation o falling within the explanatory scope of a theory T and o being an evidential support for T. Hence, our criterion for potential observational consistency is also in analogy to the one for potential explanatory power:

> The potential coherence of a cognitive system is increased by the presence of evidential support by an observation o (or a class of observations o), especially if o is significant. The evidential support provided by o is significant in case o represents an explanatory anomaly for the dominant and/or other pursued rivals.

3.5.4 Programmatic Character

So far we have presented adjusted versions of all three Bonjour's criteria, suitable for the evaluation of the initial potential coherence of a theory T. These criteria highlighted attractive features of T with respect to the criteria that constitute Bonjour's notion of coherence.

But in which way can we assure that a theory, which for instance offers certain significant explanations, but is at the same time facing some difficult explanatory anomalies, a lack of inferential connections or certain inconsistencies, is still worthy of pursuit?

First of all, it is important to notice that the fact that the theory cannot solve such anomalies at the moment, does not mean it cannot work towards finding the solutions. Especially in the early stages of a pursued theory, models are still highly idealized and outline only the main features and ideas of the young program. Thus, for instance, the empirical accuracy might in many ways still be suboptimal. Only an iterative process of gradual refinement can successively reduce the abstract character and fine-tune the models. What needs to be assured then is that there are certain ways in which the given theory can further be investigated. This is the purpose of our next criterion.

> *Programmatic character of a developing theory:* The potential coherence of a cognitive system is increased to the extent to which the system has a programmatic character and decreased to the extent it lacks one. A cognitive system has a programmatic character if it is embeded in a theoretical and methodological framework which allows for the further research of the system to proceed in spite of the encountered problems, and towards their systematic resolution.

Such a character of a cognitive system can be explicated by an example from science policy. If we are to finance a new scientific project which offers some significant explanations, but at the same time faces certain anomalies,

what we would expect from the scientists arguing for its epistemic pursuit worthiness is to show that their further research can proceed in spite of these anomalies, as well as towards their systematic resolution. In other words, a theory should exhibit an adequate perspective of the problem horizon into which it is embedded and provide the heuristics capable of guiding the research towards a possible resolution of the key problems with which the theory is currently confronted. As Kitcher points out: "to defend a particular proposal for modifying consensus practice, one must show, constructively, that it has *the potential to find solutions* to the predicaments that it faces" (Kitcher 2000, p. 31, italics added). So, even if the theory does not exhibit a high *actual* explanatory power (and/or some other virtues regarding the inferential density or consistency) at the early stages of its development, if the scientists show that they are capable of proceeding with the further research, there is a good reason to think that the project has a certain level of resilience concerning the current anomalies. And even though we cannot be absolutely certain that the research will have a successful outcome, such a programmatic character should give us a reason to make a leap of faith.[17]

For example, we have already mentioned Bohr's theory of atom as based on inconsistent assumptions. However, the theory showed not only an explanatory significance, but also a methodological plan for introducing gradually improved models (Lakatos 1978, p. 63). In that way, the programmatic character gave credibility and a *rationale* to an inconsistent program (ibid., p. 64).

Another example is Wegener's theory of the continental drift. The theory had high potential explanatory power since it provided explanations for the phenomena (such as paleontological similarities between the continental regions on the opposite sides of the ocean), for which its rivals only had weak explanations (for instance, in order to explain such similarities they had to introduce the hypothesis of land bridges between the continents, which was inconsistent with the theory of isostasy). Nevertheless, many geologists regarded Wegener's theory as incapable of accounting for its main element – the mechanism of drift. Not only was it unclear which forces could be responsible for continental drifting, but according to some, any conceivable mechanism of the drift seemed to conflict with the physical theory, since continents couldn't simply plough through the hard ocean floor ((Le Grand 1988, p. 129); (Laudan 1981, p. 230)). Nevertheless, Wegener's theory, already in its early stages had a programmatic character. Using the measurements obtained by others, Wegener estimated that the oceanic crust was no more than 5 km thick in contrast to continental blocks, the average thickness of which was taken to be around 100 km. Consequently, in Wegener's model continents moved mainly through the fluid substrate, and had only a thin semi-rigid oceanic crust in their way. In reply to the objection that such a plow would result in deformations of the ocean floor, Wegener called upon isostasy which precluded the formation of significant elevations in the seefloor (Oreskes 1999, p. 78-79). Thus, as Naomi

[17]Similarly, Whitt speaks of "programmatic research directives" that are provided by the heuristic of a theory (Whitt 1992, p. 621); also see (Whitt 1990, p. 472-473).

Oreskes explicates, even though Wegener's model may seem completely wrong from a nowadays geological perspective, in the 1920's it was consistent with the available understanding of terrestrial kinematic properties. And even though the model was based on hypotheses which could not be proved, it showed that the problem of the mechanism could in principle be resolved. Moreover, it pointed to the problem-field that required further investigation: isostasy and the nature of the substrate, which seemed to be directly related to the question of the mechanism of drifting. Thus, it gave a programmatic character to the theory of continental drift with respect to this issue.[18]

3.6 Successive Pursuit Worthiness

So far we have explicated the notion of the initial pursuit worthiness. However, as the research of a young scientific theory continues, these criteria alone cease to be sufficient for answering the question, whether the theory has remained worthy of pursuit. What we expect from a young theory in order to remain worthy of pursuit, is also to show an improvement in its epistemic properties. The successive pursuit worthiness evaluates whether a theory has remained worthy of pursuit.

We have already seen that Laudan emphasizes the link between pursuit worthiness and "the rate of progress" (see Section 3.3). Similarly, for Lakatos each step of a research programme should be consistently content-increasing, constituting thus a consistently progressive theoretical problem-shift (Lakatos 1978, p. 49). A developing research programme "is planted, as it were, in an inimical environment which, step by step, it can override and transform" (Lakatos 1978, p. 55). Such a diachronic element is rooted in the dynamic character of pursuit itself: while both the actual coherence and the potential coherence evaluated for the initial pursuit worthiness capture the epistemic state of the system at a particular time point, the potential coherence regarding the successive pursuit worthiness takes into account both the development of the system, as well as its potentials for the future endeavor. It evaluates whether the system is on its way to overcome its shortcomings and to fulfill its promising aspects that are e.g. indicated by its programmatic character. In this way we check whether we were and are still justified in giving a leap of faith to the theory and hence in taking the risk that comes with it.

Of course, we expect our theory to still fair well with respect to the criteria that were so far introduced. Hence, when we evaluate whether a theory remains worthy of pursuit we again apply our four criteria C1–C4. However, our evaluation should get increasingly more critical as time goes by in the sense that we expect the theory to live up to its promise and to exhibit growth in (at least one of) two respects:

1. *Theoretical Growth:* The scientists should make theoretic progress for instance by applying the programmatic aspects of the theory and hence

[18] See also Chapter 4.

gradually overcome problems such as explanatory anomalies, inconsistencies, etc.

2. *Growth of the Programmatic Character:* Problems that have not yet been addressed by the programmatic character of the theory should gradually become addressed by it.

Hence, these two aspects of a theoretical development function as *constraints* posed on C1–C4. A theory should either fulfill the promises that were indicated by its programmatic character and thus exhibit a theoretical growth, or at least improve the programmatic character itself. How do these two aspects influence the evaluation of pursuit worthiness in terms of our four criteria?

Let us first take a look at the theoretic growth. As time goes by we expect that progress is made at least concerning the issues that have been addressed by the programmatic character of the theory. For instance, if scientists manage to get results by making use of the heuristics provided by it then this in turn serves as a positive feedback for the programmatic character and gives us reasons to firm our leap of faith in the theory. Fulfilling the promises indicated by the programmatic character should result in a theoretical growth in one or more of the following ways:

1. the growth of the explanatory power: by introducing additional explanations or improving already existing explanations (by refining or deepening them, by introducing new evidence, etc.);
2. the growth of the internal, resp. external inferential density;
3. the growth of the internal, resp. external consistency.

In other words, the more promises (announced by the programmatic character) the theory manages to fulfill, the better it becomes with regards to its potential coherence. Moreover, if the scientists manage to have a high rate of progress, this will boost our trust in the theory even more. After all, if the heuristics and methods are demonstratively applicable in a smooth way this provides an index of the accuracy and adequacy of the theory and indicate that the theory is on its way towards satisfying our epistemic goals. Note that such a theoretical growth can also result from the improvements that were not announced by the programmatic character, for instance from parallel developments in other scientific theories (e.g., an external inconsistency with an established theory from a different domain may be resolved by the developments in that domain alone).

However, should the scientists not manage to make any progress by means of the heuristics, this will decrease its potential coherence. The reason is parallel to our argument above: if the work along the methodological and heuristic outline of the programmatic character proves to be rather tenacious or stopped entirely then this is a good reason to evaluate the programmatic character worse than initially. This will especially be the case if no alternative heuristics or methods are available, since the theory then seems to be on the wrong path. Moreover, the unresolved anomalies that were supposed to be resolved, will lower the score of the theory in the other criteria. For instance, if the programmatic character was initially meant to address certain explanatory anomalies,

the fact that this plan did not show any progress will in turn lower the potential explanatory power of the theory.

The second constraint which is relevant for the assessment of the successive pursuit worthiness is the growth of the programmatic character itself. The more time goes by the more we expect the scientists to provide theoretic and methodological means to address open questions that have not been addressed by the programmatic character thus far. This may concern the explanatory anomalies, (internal or external) inconsistencies, or the (external or internal) inferential density the theory displays.

The main underlying idea of our evaluation of the potential coherence regarding the successive pursuit worthiness is that the problems addressed by the programmatic character move to the set of resolved problems, while the (old or new) problems that have neither been resolved nor addressed by the programmatic character become addressed by it.

Let us look at an example. In the case of the pursuit of the theory of continental drift, Wegener's model, which as we have seen gave a programmatic character to his theory, was further improved. For instance, Arthur Holmes proposed a model, according to which the continents were displaced by convection currents in the earth's mantle, generated through radioactivity (Frankel 1979, p. 68). This model was inferentially connected with the theory of radioactivity, with which the rivals of the theory of continental drift were not connected. Moreover, even though Holmes' hypothesis of convection currents could not have been substantiated at the time, it pointed to a possible path towards a resolution of the problem of the mechanism of drift, and offered heuristics for its further investigation (such as the examination of the presence or indices of the presence of convection currents in the interior of the earth).[19] In addition, the number of significant explanations was increased as well through the work by Alexander du Toit, who found various similarities between the South African and the South American coastlines on the opposite sides of the ocean (Oreskes 1999, p. 157-166).

It is important to notice that the successive potential coherence depends also on some non-epistemic criteria, such as the number and the expertise of scientists working on the theory or the appropriate funding that allows for the required resources. Hence, when evaluating the successive potential coherence, and the *satisfactory* rate of growth, these factors should be taken into account as well. For instance, even though Wegener's theory of continental drift exhibited a theoretical growth throughout the 1920s and 1930s, in the 1940s hardly anyone was working on it. Hence, it is not surprising that the theory did not exhibit any significant growth in this decade.

[19]This was precisely the outcome of the expeditions conducted in the early 1930s, which included Dutch geologist Vening Meinesz and North American geologists Richard Field and Harry Hess. The results of their investigation confirmed an uneven distribution of radioactive constituents and thermal properties in the earth, which made Meinesz conclude that "in the actual earth there can be no doubt that convection currents must develop" (quoted from (Oreskes 1999, p. 248)).

3.7 Meta-Justification

In this section we will give a meta-justification of our criteria for the pursuit worthiness of a theory. The goal is to argue that following the lead of our criteria is indeed conducive of epistemic aims of science. In Section 3.1 we have specified two major epistemic goals of science: (a) to gain adequate and accurate knowledge about the world and (b) scientific knowledge should be robust, i.e. the scientific knowledge base as a whole should be able to avoid and/or withstand perturbations. One may view science as a long-winded dialogue with nature with the final aim to come to an agreement.[20] Robustness concerns the situation when nature answers "No" to our arguments: in these cases we want to be able to respond. Sometimes it is enough to change our arguments viz. theories, but sometimes – if nature's "No" is too plentiful – it is more promising to come up with new argumentative strategies viz. theories and to give up on the old one. The central question in this chapter is: when are we justified to pursue a new cognitive system T judged by its epistemic virtues? In view of our dialogue with nature this is so if T is promising of strengthening our repertoire of argumentative moves to dodge nature's "No"'s. More concretely, in view of our epistemic goals and the coherentist perspective adopted in our account this is the case if T is *promising* of growing into a theory that is (a) highly actually coherent and (b) that strengthens the robustness of our scientific knowledge as a whole. It is promising of the latter in case it is promising of growing into a backup candidate in case the dominant rival falls into crisis. Of course, since T is in its development or yet to be elaborated further we can only talk about its epistemic promise or potentials. This comes with an essential risk and it is a worthwhile goal to control and reduce the latter as much as possible. Hence, we are interested in indices of the epistemic promise of theories. The epistemic justification in the context of pursuit can then be carried out by evaluating and comparing theories with respect to these indices.

Two indices of epistemic promise are central in our approach: (i) the focus on *significant aspects* concerning criteria of coherence and (ii) the *programmatic character* and the focus on *developmental aspects* that comes with it.

Let us first discuss point (i). Take for instance the potential explanatory power (see Section 3.5.1). Here the focus was on *significant explananda*. On the one hand, we have pointed out crucial and benchmark problems. In order for a theory to be epistemically attractive, i.e. to provide an adequate grip on its subject matter, it should be able to tackle these problems (or at least provide a perspective for tackling them, see (ii)). On the other hand, some of these are phenomena for which the dominant rival does not offer a (satisfactory) explanation. In view of the robustness of scientific knowledge we are interested in tackling these problems. After all, these (and other) anomalies may very well be deficits of the dominant rival that may eventually be a part of the factors

[20] Pera rightly pointed out that it is more accurate and awarding to view science as a discourse between multiple agents and nature (Pera 1994) (see Chapter 5). However, in order to make our point it is fine to simplify matters.

responsible for its downfall.[21] In this case we need backup theories that can fill this lacuna and offer alternative argumentative strategies to the "No"'s of nature that produced them.

A similar situation occurs with respect to consistency and inferential density. In case the dominant rival is inconsistent with another established theory, or it doesn't offer inferential connections with it, this again may be an index of its deficiency. Hence, in order to gain a robust scientific knowledge base we are again interested in theories that are able to fill these gaps.

Since we are evaluating the pursuit worthiness of theories it is natural to also address developmental aspects. This brings us to point (ii). Theories in the context of pursuit are suboptimal in various respects: they are burdened with explanatory anomalies, they may have to deal with inconsistencies, etc. Hence, the only way to compensate for these shortcomings is to offer a programmatic character, i.e. a methodological heuristics for how to proceed further and how to eventually overcome (some of) the shortcomings. The programmatic character indicates that a theory has the potential to grow into a richer and more (actually) coherent one.

Of course, as time goes by we expect from a theory to actualize these promises given by its programmatic character and also to widen the programmatic character to address formerly open problems. These aspects are taken into account in our evaluation of the successive potential coherence (see Section 3.6). Our criteria require from a theory to exhibit a theoretical growth or at least to develop its programmatic character. In this way we make sure that the unresolved anomalies (that is, explanatory anomalies, the lack of a unifying core of hypothesis, the lack of inferential connections with other established theories, the lack of internal consistency, or the lack of consistency with other established theories) are either resolved by the theory or addressed by its programmatic character. Moreover, they make sure that as time goes by, these anomalies become more urgent to be resolved, and that the epistemic status of the theory is developing, or shows promising of developing, towards the actual coherence. At the same time, such a development contributes to the robustness of scientific knowledge: the more coherent the theory is, the more entrenched it is, and so, the more robust our knowledge becomes.

3.8 Conclusion

In this Chapter we have presented an account of epistemic justification suitable for the evaluation of theories in the context of pursuit. We have built our model paradigmatically on the basis of Bonjour's coherence theory of epistemic justification. By adjusting the criteria of his concept of coherence and by complementing them with the requirement for a programmatic character we gave an account of what we have dubbed *the potential coherence* of a cognitive

[21] There may of course be more than one dominant rival. In this case our discussion can easily be adjusted accordingly.

system. This enabled us to judge whether the theory is sufficiently promising to be further investigated.

In this conclusion we want to point out a specific contextual character of our framework. Many notions constituting our criteria have been left undefined in any strict manner. For example, what counts as a scientific explanation, what counts as a benchmark problem in the field, what counts as a unified core of hypotheses, etc. – have been intentionally left unspecified. The meaning of these notions needs to be obtained in view of the specific conceptual framework characterizing the cognitive horizon of the time. However, such an approach introduces the problem of relativism: if the criteria of potential coherence can be interpreted in different ways, doesn't that mean that a theory may be evaluated as worthy of pursuit in view of one interpretation, and not worthy of pursuit in view of another? How can we avoid such a relativistic outcome in cases of scientific controversies, in which, for example, the new theory may challenge the traditional views on what counts as a valid standard of scientific explanation or what counts as a valid explanandum in the given field? A detailed analysis of this problem would have to take into consideration not only the change in the interpretation of the standards and in the weighting that are placed on them, but also the change in more general goals of science. We will say a bit more about this problem in Chapter 5. At this point, it is important to mention that our account may not be sufficient for the modeling of radical shifts in epistemic standards, in which the very criteria constituting it may have to be altered.

It is also important to point out that we have built our account *paradigmatically* on the basis of Bonjour's coherentism. Bonjour's criteria may be suboptimal in certain respects when applied to the evaluation of scientific theories,[22] but they were sufficiently simple for demonstrating our approach. Our main aim was to show how the standards of epistemic justification suitable for the context of acceptance can be modified and adapted for the context of pursuit. Indeed, by modifying Bonjour's account or by replacing it with some other set of criteria, we could in an analogous way obtain alternative frameworks of epistemic justification, suitable for the evaluation of pursuit worthiness.

[22]For instance, it could be argued that the criterion of predictive power is insufficiently represented in Bonjour's account, or that the virtue of robustness of theories – although indirectly addressed in terms of inferential density – could be presented in a more elaborated way (see footnotes 1 and 5).

Chapter

4 Rationality and Irrationality in the History of Continental Drift

Was the Hypothesis of Continental Drift Worthy of Pursuit?

✎ *This chapter is based on a paper with the same title. It is co-authored by the second author Erik Weber. We wish to thank Steffen Ducheyne and Christian Straßer for the valuable comments on a former draft of this paper.*

Summary The revolution in geology, initiated with Alfred Wegener's theory of continental drift, has been the subject of many philosophical discussions aiming at resolving the problem of rationality underlying this historical episode. Even though the debate included analyses in terms of scientific methodology, applications of concrete accounts of epistemic justification to this case study have been rare. In particular, the question as to whether Wegener's theory was epistemically worthy of pursuit in the first half of the twentieth century, that is, in its early development, remained open or inadequately addressed. The aim of this chapter is to offer an answer to this question. The evaluation of Drift will be done by means of an account of theory evaluation suitable for the context of pursuit, developed in Chapter 3. We will argue that pursuing the theory of continental drift was rational, i.e., that it was irrational to reject its pursuit as unworthy.

4.1 Introduction

Ever since the revolution in the earth sciences culminated in the overall acceptance of the theory of plate tectonics, philosophers and historians of science have been analyzing this shift in geology. What made the development in geology very interesting is its specific dynamics. Even though the hypothesis of continental drift was proposed by Alfred Wegener already around the 1920s, it was firmly rejected by many geologists not only as unacceptable but even as

unworthy of further pursuit. Almost half a century had to pass in order for this hypothesis to become finally accepted and elaborated into the theory of plate tectonics. Such a development inspired discussions among philosophers and historians of science, which especially focused around two issues: first, the nature of the revolution in geology and the applicability of different methodological frameworks to it, and second, the rationality or irrationality of the stances of scientists throughout the revolution.

With regard to the first question, some philosophers and historians of science argued that this episode can be described in terms of Kuhn's notion of scientific revolution (e.g. (Stewart 1990)). However, the majority of them agreed that Laudan's account of progress of science is more suitable for this case-study (see e.g. (Le Grand 1988), (Frankel 1979), (Laudan 1987)). The aptness of Laudan's framework stems from two important notions in his account. On the one hand, his notion of a research tradition as a broader theoretical framework, constituted by specific scientific theories, has been shown useful for capturing the rivaling camps in geology, neither of which could be reduced to one generally accepted theory.[1] On the other hand, his distinction between the context of pursuit and the context of acceptance (Laudan 1977, p. 108-110) proved to be important for analyzing questions of rationality regarding this case study. According to Laudan, "acceptance, rejection, pursuit and non-pursuit constitute the major cognitive stances which scientists can legitimately take towards research traditions (and their constituent theories)" (Laudan 1977, p. 119). As the names suggest, the context of acceptance deals with the question as to whether a certain theory is to be accepted as the standard in the given field, while the context of pursuit deals with the question as to whether a (possibly young, undeveloped) theory is at all worthy of further pursuit. Such a distinction allows for a twofold analysis of the rationality of judgments made by geologists during this time. This brings us to the second important topic that attracted interests of philosophers and historians of science, as we have mentioned above.

With respect to the context of acceptance, there has been a general agreement that it was rational to reject Wegener's theory of drift when it first appeared in the 1912-1915, and to accept Drift[2] in the early 1960s, after it had developed into the theory of plate tectonics. With respect to the context of pursuit, most of the authors simply described the views of geologists at the time. Nevertheless, the question: *Was pursuing the theory of continental drift*

[1] In contrast to Lakatos' notion of a research program, Laudan's notion of research tradition allows for an evolving hardcore, which, according to some authors, makes it especially suitable for describing the geological sciences in the first half of the twentieth century (see e.g. (Frankel 1979, p. 53)). According to Laudan, what makes theories belong to the same research tradition is its hardcore, which, even though sacrosanct for its proponents, can still evolve. In other words, theories belong to the same tradition not because some of their crucial assumptions are identical, but rather because these assumptions overlap (see (Laudan 1977), p. 99).

[2] As it is usual in the literature on this case study, we will call this research tradition and its constituting theories (running from Wegener's theory of continental drift to the theory of plate tectonics) – Drift.

in the first half of the twentieth century at all rational (in the sense of being epistemically justified), or was it rational to reject its pursuit? – so far has not been properly addressed.

The aim of this chapter is to answer this question in terms of a concrete account of epistemic justification. To this end, we will use the account developed in Chapter 3, which will be slightly modified in order to allow for the notion of pursuit worthiness in a strong sense.

The chapter is structured as follows. We begin in Section 4.2 by offering a brief historical overview of the revolution in geology. In Section 4.3 we will show that so far this question has not been properly addressed in the literature on this historical episode. In Section 4.4 we will present the account of epistemic justification which we shall employ in this chapter. Sections 4.5 to 4.8 bring an evaluation of pursuit worthiness of Drift in the first half of the twentieth century. In Section 4.9 we will discuss the epistemic stances of geologists in this time period in view of our evaluation. Moreover, we will point out the importance of the evaluation of epistemic pursuit worthiness and some undesirable implications for scientific debates once this type of assessment has been neglected. Section 4.10 concludes this chapter.

4.2 Historical Overview of the Revolution in Geology

4.2.1 Rivaling Theories

Alfred Wegener launched the idea of continental drift in a short article in (Wegener 1912). A more elaborate version appeared as a book in 1915 (Wegener 1915). His central claims were that all the continents had once been united, had broken apart and had drifted through the ocean floor to their current locations (Le Grand 1988, p. 1). Following Le Grand we call adherents of the of large scale lateral movements of continents drifters. Wegener probably was not the first drifter, but he was the first to develop a full-fledged argumentation for it. He tried to show that Drift is superior to the two theories that already existed, viz. permanentism and contractionism. The summary of views and arguments presented below is based on (Le Grand 1988) (especially p. 19-28, 40-46 and 55-57).[3]

According to permanentists "... the continents were formed in remote geological times as the earth had gradually cooled down and contracted. Since then, they had been permanent features of the earths surface." (Ibid., p. 20-21). Continents do not move laterally (this is how permanentism differs from Drift) and do not disappear (this is how it differs from contractionism, see below). James D. Dana, professor at Yale University from 1850 to 1892 and one of the main adherents of this theory, used the slogan "Once a continent, always a continent; once an ocean, always an ocean" (quoted from: Ibid., p. 21). According

[3]Even though Le Grand offers a good overview of this debate, some details that are relevant for our discussion are better worked out in (Oreskes 1999) (see Sections 4.5-4.8).

to permanentists there have been small elevations (producing mountains) and small subsidences (producing e.g. shallow inland seas).

The most important representative of contractionism was the German geologist Eduard Suess. The central claim of his theory was the following: "As the earth lost its heat, a rigid crust formed. As the earth continued to cool and shrink, this crust wrinkled, folded and subsided" (Ibid., p. 25). This process explains how mountains come into existence (lateral movement of parts of crust). The collapses occur sporadically, creating new oceans and new continents: when the crust collapses in a certain region, the water flows to the new lowest point; continents thus can become oceans and oceans land. Because the collapses occur sporadically, contractionists see the history of the earth as divided into periods of rapid change and periods of stability.

According to Wegener the earth consists of a series of concentric shells with different compositions and densities (highest density in the near to the core). The temperature also is higher closer to the core. The continents constitute the outermost shell and consist of blocks of *sial* (silica + alumina) which (like icebergs in the sea) partially float on and extend into blocks of *sima* (silica + magnesia). The oceans are situated between the blocks of sial, and ocean floor is made of sima. The continents once were united in the super-continent Pangaea, which broke apart in the Cretaceous (Wegener did not give e reason for this break). Since then, the continents are propelled by one or more forces through the ocean floor. Sometimes they move apart (Africa and South America). Sometimes they collide, resulting in mountains (e.g. the collision of India and Asia creating the Himalayas).

In order to understand Wegener's theory and the arguments discussed below well, it is important to elaborate the analogy with the icebergs. Icebergs are solid, while the water in which they float is fluid. Analogously, the continents are solid, and the sima in which they float is (relatively) fluid. The density of the icebergs is lower than that of the surrounding water, so they float; analogously, the density of the continents (sial) is lower than that of the surrounding ocean floor (sima), so the continents float. Wegener invoked two forces: pole-flight (Polflucht, a force due to the rotation of the earth and directed form the poles to the equator) and a tidal force (from east to west) as a result of gravitational attraction of the sun and the moon.

4.2.2 Arguments in the Debate

Here is an overview of the arguments that were exchanged:

1. Permanentism cannot explain the distribution of fossils (palaeobiogeography) and living species. If oceans and continents are permanent, similarities between species separated by oceans cannot be explained. Drift explains the similarities, because the continents were once united. contractionists assumed that there are sunken continents and/or sunken landbridges that connected the continents we have now. These sunken continents and landbridges explain the similarities. Suess postulated the

existence of the palaeocontinent Gondwana, of which the central part sunk in the Indian Ocean. What remained is now: Australia, India and Africa. For the connection between Africa and South America he left the two options open. In the terminology of (Laudan 1977), the distribution of fossils and living species is an anomaly for permanentism.

2. Permanentism cannot explain geological similarities between continents. Mountain chains and coal-basins seem to continue on both sides of the Atlantic Ocean (e.g. coal fields in Pennsylvania on one side, French-Belgian coal basin on the other side). Drifters can easily explain these similarities. Contractionists have to assume that the sunken parts have the same geological structure as the parts to be explained (so they formed a geological connection, not just a route for plants and animals). Permanentists have to claim that the similarities are accidental. This is another anomaly in Laudan's terms.

3. Contractionism is incompatible with isostasy (see above: sima and sial) that was well supported at the beginning of the 20th century by all kinds of measurement. If the landbridges consist of sial, they cannot sink through the denser sima of the ocean floors. The same for the sunken connecting continents. In Laudan's terminology, this is an external conceptual problem for contractionism.

4. Contractionism is incompatible with the presence of radioactive materials in the earth's crust. Physicists discovered that radioactive material was widely distributed in the earth's crust, and that they produce heat when decaying. In 1909 the Irish physicist John Joly argued that for this reason it is very problematic to maintain that the earth cools down through loss of internal heat. Rather, the most plausible state is that the temperature of the earth remains constant or increases a little bit. This is incompatible with the extreme cooling down that contractionists have to assume (e.g. 1200°C to explain the formation of the Alps, much more for higher mountains). This is another external conceptual problem for contractionists.

5. As mentioned above, Wegener invoked two forces (pole-flight and tidal force) that propelled the continents. The problem is that these forces are too weak:

> The earth did behave like a fluid in some respects, but no one was proposing that the ocean floors were in fact liquid: they were composed of dense, basaltic rocks. How could the continents move laterally through such floors without crumbling to bits? What enormous force not only moved the continents but had crumpled them up to form the Alps, Rockies, Andes and Himalayas? The forces which Wegener invoked did exist but they were far too weak. A force nearly 1 000 000 times

> stronger was needed, and if it did exist it would surely have been noticed by physicists. (Le Grand 1988, p. 55-56)

In Laudan's terminology, this is an external conceptual problem. The problem is of a less strong type than the conceptual problems facing contractionism: physics makes Drift implausible, but there is no logical incompatibility.

6. The complementary shapes of coast lines (e.g. of Africa and South America) can be explained by Wegener if he assumes arbitrary changes of shape. His rivals judged that he could not explain the jig-saw fit, and that this was a problem for all three theories. In Laudan's terminology, this is an unsolved problem for all three theories.

7. Wegener tried to develop geodetic evidence. He participated in several expeditions in Greenland (and died during one of them in 1930). He tried to measure whether Greenland was moving. His results were well within the margin of error of his apparatus and method (it was not sure that the measurements were made at exactly the same spot). In Laudan's terminology this is a failed attempt of Wegener to create an extra anomaly for his rivals.

4.2.3 Seafloor Spreading

Given these arguments for and against the three theories, it is not surprising that none of them became dominant. Each of them was confronted with a number of problems. The situation changed in the 1960s, after findings in paleomagnetic studies and investigations of the seafloor gave additional evidence for Drift. Moreover, Harry Hess' idea of seafloor spreading offered a mechanism of drift that both solved the problem of mechanism and was supported by a sufficient amount of evidence:

> The nub of his theory was that new seafloor was generated at ridges by the upwelling of mantle material. The old seafloor gradually moved from the ridges and was eventually dragged down at the trenches and reconverted into mantle. The cycle was driven by convection currents rising under ridges and descending under trenches.
> ...
> The continents were passive passengers seated on dynamic ocean floors. Hess's model provided a solution to the conceptual problem with which Drifters have wrestled for fifty years: how could the continents drift through the ocean floor? Hess's answer was that they moved with the crust, not through it. (Le Grand 1988, p. 197)

In order to see why the conceptual problem disappears, it is important to notice that the new theory requires different forces (at other places, and less

strong).[4] After Hess introduced this idea, Drift became the dominant theory very rapidly. Adherents of the other theories changed side.[5]

4.3 Others on Drift in the Context of Pursuit

In this section we will discuss the work of other scholars that have analyzed the development of Drift in the first half of the twentieth century. Our aim is to show that the question as to whether Drift was worthy of pursuit has been either neglected or inadequately addressed.

On the one hand, most of the authors who analyzed this revolution focused primarily on the context of acceptance. For example, Paul Thagard (in his (Thagard 1992)) offered an analysis of this revolution in terms of his account of explanatory coherence. His primary aim was to answer the question: *"Why was Wegener's theory of continental drift largely rejected in the 1920s, and why, in contrast, were the new ideas about seafloor spreading and plate tectonics largely accepted in the 1960s?*" (p. 171, italics in original[6]). Both research questions address only the context of acceptance. Similarly, Ronald Giere in his (Giere 1988) discussed why there was no revolution in the 1920s in contrast to the 1960s (p. 227-277).

On the other hand, those authors who did discuss the rationality in the context of pursuit, did not take into consideration what we are primarily interested in. They were concerned with the question as to why some geologists pursued Drift and why some others did not pursue it, but not whether pursuing Drift in general was rational or not.[7] For example, Miriam Solomon in her (Solomon 2001) explicates which epistemic and non-epistemic factors influenced a dissent among geologists in the first half of the twentieth century, and which factors later on led to the consensus over the theory of plate tectonics. Her approach aims at explicating why some geologists decided to pursue Drift, and why some others decided not to purse it. But this does not answer the question: Was Drift around the 1920s-1930s epistemically worthy of pursuit or not?

Another example is Henry Frankel who agued in his (Frankel 1979) that if Laudan's methodological framework is applicable to the development and reception of Drift, then this tradition "should not be accepted by the relevant scientific community if it does not have greater problem-solving effectiveness than competing traditions, and *should be pursued only by those scientists who believe it has promise of its future ability to solve problems*" (p. 75; italics added). In addition, he suggested that for certain geologists (such as Arthur

[4] As we shall see, Arthur Holmes proposed a similar mechanism of the drift already few decades earlier. For a detailed discussion of Hess' proposal and its development see (Frankel 1980).

[5] Some geologists accepted Drift only after the evidence from some of its novel predictions was confirmed (see (Laudan and Laudan 1989, p. 221)).

[6] Henceforth, italics appearing in a quoted text are present in the original as well, unless otherwise indicated.

[7] Most of these approaches make use of Laudan's criterion for when pursuing a theory is rational. We will take a closer look at Laudan's criterion in Section 4.4.

Holmes and Alexander du Toit) the theory of drift was sufficiently promising since it could solve a number of empirical problems (in Laudan's sense of the term) without necessarily creating more conceptual or empirical problems. Without going into discussion on the notion of rationality underlying such a "subjective" assessment of pursuit worthiness, let us just notice that Frankel's approach to the rationality in the context of pursuit does not answer the question we are interested in. What we wish to investigate is not whether Drift was worthy of pursuit *for some*, but whether it was worthy of pursuit in general, for the geological community at the time.

In a similar vein, Homer Le Grand (Le Grand 1988, p. 95) as well as Rachel Laudan (Laudan 1987, p. 205-213) both explicate why some geologists engaged in the actual pursuit of Drift, and why other geologists did not pursue it. As mentioned above, such an approach does not answer our research question.

Let us then present the account of epistemic justification in the context of pursuit, which we shall use to evaluate Drift.

4.4 The Notion of Pursuit Worthiness in a Strong Sense and the Main Claims of the Chapter

As we have announced, we will show that Drift was *worthy of pursuit in a strong sense* (henceforth: WPSS).[8] By this we mean that it satisfied the criteria of pursuit worthiness in the most important respects. More precisely, our main claims are, first, *that Drift was initially WPSS*, and second, *that it remained WPSS* throughout the first half of the twentieth century.

We say that a research tradition is initially WPSS if it satisfies the following criteria, formulated on the basis of our account of potential coherence:

1. *Presence of Significant Explanations:* A theory has to offer explanations that are significant at that point of the scientific development.[9]

2. *Inferential Density:*

    a) *Internal Inferential Density:* A theory should not have a less unified core of hypotheses than its dominant rival or it should be able to address such a lack by means of its programmatic character.

    b) *External inferential density:* A theory should be at least as inferentially connected with established theories from other scientific domains as its dominant rival or it should be able to address the lack of such connections by means of its programmatic character.

[8] We will use the same shortcut also for "pursuit worthiness in a strong sense".
[9] For what counts as a significant explanation here see Chapter 3.

3. *Programmatic character:* A theory should have a programmatic character that addresses all the major problems of the theory (such as explanatory anomalies, inconsistencies, etc.).[10]

We say that the research tradition remains WPSS if the previous criteria are satisfied in view of:

1. a *theoretical growth* or

2. a *growth in the programmatic character*.

In contrast to our account of potential coherence, which allows for a gradual evaluation of pursuit worthiness of theories (or research traditions), the criteria of WPSS are presented in a discrete manner and (to make things simpler) without introducing the concept of coherence. In other words, the criteria of WPSS are formulated in such a way that if a theory or a research tradition satisfies them, we can say that it is (or in case of the past theories: was) certainly WPSS.

It is important to notice that satisfying the criteria of WPSS is a sufficient condition for a research tradition to be worthy of pursuit, even though it may not be a necessary one.[11] There are various examples of scientific theories that were worthy of pursuit even though they did not fully satisfy all of these criteria.[12] In such cases, a theory has to score very high in some respects in order to compensate for a low score in others. However, by showing that Drift was WPSS we can avoid a discussion involving such a weighting of its properties. Moreover, showing that Drift was WPSS makes it easier to argue that the view according to which Drift was not at all worthy of pursuit can be considered epistemically unjustified.

When evaluating pursuit worthiness of research traditions, we are in fact evaluating the pursuit worthiness of its constituting theories. Therefore, examining the initial WPSS of Drift will refer to Wegener's model. However, it is important to mention that even though Wegener's theory appeared around 1915, most North American geologists became familiar with his work only in 1924, when English translation of Wegener's book was published (Marvin 2001, p. 21). This complicates rooting the initial WPSS of Drift in a specific year, and somewhat blurs the demarcation line between the initial WPSS and the property of remaining WPSS. Our approach will be the following: we will answer the first issue in view of the discussion that occurred in between 1912 and the mid-1920s, including Wegener's model of Drift as well as the first criticisms

[10]The criterion of consistency has been skipped since for the notion of WPSS we are focusing only on possible problems a theory may have in this respect, i.e. inconsistencies, for which we require to be addressed by the programmatic character.

[11]That means that if a theory satisfies our criteria then it is (or was) WPSS; however, if it does not satisfy them, we cannot make any claims about its WPSS.

[12]This is especially the case with theories that are initially internally inconsistent. For example, a theory with a strong explanatory power, whose programmatic character might not address some inconsistencies in it, could still be worthy of pursuit (see (Lakatos 1978, p. 55)).

of it. The latter issue will be discussed in view of the arguments and alternative models of Drift that were offered by Wegener's followers (even though some of them were proposed already in the early 1920s).

The historical presentation will be mainly based on (Oreskes 1999), which is one of the most recent studies of this episode, as well as on (Le Grand 1988).

4.5 Presence of Significant Explanations

As we have mentioned in Section 4.2, when Wegener first proposed his theory in 1912 its main explanatory rival – Suess's theory of contraction – had been under a serious attack (also see (Oreskes 1999, p. 21-55)). If the principle of isostasy was correct (and by the time of Wegener's proposal there was a growing conviction that it was), continents and ocean floors had to be different either in structure or in composition. This conflicted with Suessian idea of interchangeability of oceans and continents, and thus with the hypothesis of land bridges, used for explaining the similarities between the coastal regions on the opposite sides of oceans. Even though some geologists found a solution for the conflict between contractionism and isostasy in the permanentist perspective, permanentism had even more trouble explaining such similarities. Therefore, both rivals of Drift – contractionism and permanentism – can in this case be regarded as offering either a weak explanation for the similarities between the continents or no explanation at all. Two particularly significant types of similarities that Wegener's theory could account for were the following.

First, the paleontological similarities. The evidence for them was at this point already well established by paleontologists independently of Wegener's hypothesis (Oreskes 1999, p. 56). Not only was there an overall resemblance of fossil forms indicating that continents must have been somehow connected in the past, but there was also an evidence of a distribution of certain organisms, such as earthworms, which aren't capable of swimming or flying or having resilient seeds or a dormant life cycle or free-floating larval stage, which could allow for their passive distribution (Ibid., p. 57). A similar case are certain species of snails that were, just like the earthworms, unlikely to have crossed all the way across the land bridges (Le Grand 1988, p. 43).

Second, there was an evidence of nearly identical stratigraphic sequences and structural patterns on the coastlines of the matching continents. Moreover, Caledonian fold belts in North America matched with the Appalachian ones in Europe, while the Gondwana beds in India were nearly identical to the Karroo sequence of southern Africa (Oreskes 1999, p. 57).

Another significant explanation that Drift offered regarded a paleoclimatic evidence. By the twentieth century there was a consensus that the earth's climate had undergone repeated fluctuations. For example, glacial deposits in South Africa and Australia indicated that the climate had been much colder in Permian. However, the cause of these climatic fluctuations, which came to be known as the problem of Permo-Carboniferous glaciation, was unknown (Ibid., p. 58). The main problem for permanentist and contractionist accounts was

that different climatic conditions occurred at about the same time at different parts of the globe, which prohibited explanations stating that the earth as a whole was once hotter or cooler than now (Le Grand 1988, p. 44). In contrast, Wegener offered an explanation in terms of a shift of geophysical poles (so-called polar wandering), as well a shift of continents relative to the poles.

Without going further into details of the explananda which the Drift was able to account for, it is important to notice that the presence of significant explanations does not mean that Drift had no explanatory anomalies. For example, the evidence of Carboniferous glaciation was also found near Boston, which, according to Wegener's account, must have been in the tropical climate at the time (Le Grand 1988, p. 56). Nevertheless, such explanatory anomalies are a usual component of young scientific theories, and should, thus, not be the reason for rejecting their pursuit. One of the main tasks of further developments of the theory is exactly to remove such problems. And that a research towards fulfilling this task can proceed is guaranteed by the programmatic character of the theory, which is our next criterion of evaluation.

We can thus conclude that Wegener's theory had significant explanations in the the sense defined in Chapter 3.

4.6 Inferential Density

When it comes to internal inferential density, it is easy to see that none of the three rivaling theories was especially unified. Drift could not provide a precise mechanism of drifting, but it was able to account for many geological phenomena with the same hypothesis (namely, the hypothesis of the continental drift). In contrast, both contractionism and permanentism had to introduce an additional hypothesis, such as the idea of land bridges and isthmian links which in the past connected the continents, in order to account for the vast evidence of similarities between the coastal regions on the opposite sides of oceans (or to leave the phenomena unexplained). Thus, Wegener's theory at least did not have a less unified core of hypotheses than its rivals.

As for the external inferential density, we will show in Section 4.7 that Drift successfully addressed potential problems with physics and seismology. Moreover, we will show that it had inferential links with the theory of isostasy, in contrast to contractionism that was not well connected with it. In view of these insights it will be clear that Wegener's theory satisfied the criterion of external inferential density as well.

4.7 Programmatic Character

Let us start with an example of how an anomaly is addressed by the programmatic character – the above mentioned evidence of glaciation in the North America. With respect to this explanatory anomaly Wegener pointed out that the vast majority of other evidence indicated that the area was in a tropical climate which made the glacial origin of these deposits suspicious. He thus

suggested that the deposits could have originated in some other way (Brooks 1949, p. 232). For instance, the Appalachian orogenic belt, in which some of the deposits were found, could have been a high mountain range at the time (see also (Holmes 1944, p. 502). This hypothesis also indicated in which direction could the further examination of this problem proceed.

In this section we will focus on two major problems Drift was confronted with: first, the problem of the mechanism of Drift, and second, the conflict between Drift and seismology. We will show that with respect to both of them Wegener's theory had a programmatic character.[13]

4.7.1 The Mechanism of Drift

As we have already mentioned, the main problem of Drift was the question of the mechanism governing the continental drifting. It has become common in the literature on Drift to argue that in the first half of the twentieth century it was not only unclear which forces could be responsible for continental drifting, but that any conceivable mechanism of the drift seemed to conflict with the physical theory (e.g. (Le Grand 1988, p. 129), or (Laudan 1981, p. 230); also see Section 4.2.2 in this chapter).

In contrast to such a view, Oreskes argues that Wegener's proposal was well rooted in the theory of isostasy. She starts off by pointing out that for Wegener the principle of isostasy was "nothing more than hydrostatic equilibrium according to Archimedes' principle, whereby the weight of the immersed body is equal to that of the fluid displaced." (quoted from (Oreskes 1999, p. 65)). Thus, continents could be seen as floating in hydrostatic equilibrium, which means that the substrate in which they are embedded has to behave, *over geological time*, in a fluid manner.

But such an idea of a mobile substrate was not a novelty of Wegener's theory. The basic idea of isostasy refers to a condition to which the crust and the mantle tend, in the absence of disturbing forces. The first conceptions of isostasy from the second half of the nineteenth century conceived of crust as floating on the denser underlying mantle (Watts 2001, p. 1). Only by introducing the idea of a fluid or plastic substrate could they account for the oscillations of the earth's crust. Airy's model of isostasy, which was well accepted throughout Europe, hypothesized that a thin layer of crust overlays a fluid layer of greater density just like timber blocks float on water (Ibid., p. 12, 21). English geologist Reverend O. Fisher in his book *Physics of the Earth's Crust* from 1881 suggested that the crust is analogous to the broken-up area of

[13]We have chosen these two problems as the major attacks on Drift on the basis of Oreskes' study. According to Henry Frankel though, the problem of explaining the Southern Glaciation was another heated topic in debates over the continental drift (Frankel 1987). Frankel presents a number of objections raised against Wegener's explanation of Permo-Carboniferous Ice Cap, but also replies by drifters, where particularly important were the arguments given by Alexander du Toit, Arthur Holmes and George C. Simpson (Ibid., p. 212-216). These replies were the basis of the programmatic character of Drift with respect to this issue, though for the details of this discussion, we are referring the interested reader to Frankel's article.

ice, floating upon water, obeying Archimedes' principle (Ibid., p. 15). Similarly, North American geologist Clarence E. Dutton spoke of the flotation of the crust upon a liquid or highly plastic substratum ((Ibid., p. 16-17); (Oreskes 1999, p. 67)). By the time Wegener's work was translated into English in 1924, there was a rough consensus among European and North American geologists that there was a mobile layer beneath the earth's crust (Oreskes 1999, p. 66-80). Moreover:

> The idea of moving continents was perhaps not as great a conceptual leap as might otherwise appear. Chamberlin, Dutton, Hayford, and others had written explicitly of 'lateral creep' and continental 'spreading'; the unknown issue was the *scale* of these effects, and whether they operated in a cyclical manner, as Chamberlin seemed to suggest, or whether they could actually produce a net lateral motion, as Hayford might be interpreted to imply. (Ibid., p. 72-73)

Oreskes adds that the prominence of the advocates of isostasy makes it unlikely that other geologists did not know about their research. And even if they were ignorant of isostasy, the idea of a mobile substrate was inherent in the theory of geosynclines which was well known among the North American geologists (Ibid., p. 73-74). Finally, the Fennoscandian uplift represented a phenomenological evidence for the mobile substrate. The uplift was a result of the removal of glacial ice, and according to Fisher, a direct consequence of isostasy. For Wegener, the Fennoscandian uplift demonstrated that the substrate had to be sufficiently mobile to flow out of the way of the depressing continent, and upon the removal of the glacial load to flow back under the continent (Ibid., p. 76).

Thus, the main novelty of Wegener's theory was not the possibility of the horizontal movements, but their scale and extent. But what about the rigid ocean floor through which, according to his model, continents had to plow? Oreskes writes:

> Wegener's argument hinged on the belief that the ocean floor was more like the crustal substrate than the continental blocks – or, to use the terminology of the day, it was simatic (rich in silicon and magnesium) rather than sialic (rich in silicon and aluminium). This was not a particularly controversial view: it had long been suggested by evidence from ocean dredging and the basaltic composition of most ocean islands. ... And if ocean floor *was* primarily composed of basalt, then ... the continents had deep roots *and* the ocean basins were composed of denser material than the continents. If so, then the continents plowed mostly through plastic substrate and needed only to dislodge a thin veneer of crust at the top. (Ibid., p. 77)

Using the results of gravity work obtained by others, Wegener estimated that the oceanic crust was no more than 5 km thick in contrast to continen-

tal blocks, the average thickness of which was taken to be around 100 km. Consequently, in Wegener's model continents moved mainly through the fluid substrate, and had only a thin semi-rigid oceanic crust in their way. In reply to the objection that such a plow would result in deformations of the ocean floor, Wegener called upon isostasy which precluded the formation of significant elevations in the seefloor (Ibid., p. 78-79).

Thus, as Oreskes explicates, even though Wegener's model may seem completely wrong from the nowadays geological perspective, in the 1920's it was consistent with the available understanding of terrestrial kinematic properties. And even though the model was based on hypotheses which could not be proved, it showed that the problem of the mechanism could in principle be resolved. Moreover, it pointed to the problem-field that required further investigation: isostasy and the nature of the substrate, which seemed to be directly related to the question of the mechanism of drifting. Thus, it gave a programmatic character to Drift with respect to this issue.

As for the origin of forces governing the drift, Wegener hypothesized two possible causes: pole-flight force and tidal retardation. Even though they turned out to be too weak to account for drift, at the time when Wegener proposed them, they contributed to the programmatic character of Drift, by indicating in which direction a further investigation of this question could go (namely, examining if these forces are strong enough to move the continents).

4.7.2 The Conflict with Seismology

The second major objection to Drift concerns the inconsistency (and hence also the lack of inferential connections) between Drift and seismology.

Harold Jeffreys was a strong opponent of Drift coming from the contractionist side. One of his main arguments against the drifters was that their theory was inconsistent with seismology (in contrast to contractionism which was compatible with it). According to him, the propagation of seismic waves at depth of the earth's interior implied a solid and rigid earth. This conflicted with the idea of a fluid substrate upheld by the drifters (Oreskes 1999, p. 83).[14]

However, Jeffreys' arguments were rebutted, on the one hand, by the theory of isostasy which required a fluid substrate, and on the other hand, by the drifters themselves. As for Wegener's reply to this objection, he argued that earth materials may behave in a rigid manner in response to short-duration disturbances, such as seismic waves. But the same materials may exhibit plasticity in response to a small, steady and slow pressure over geological times (Ibid., p. 79). Such a reply showed that further examination of the substrate

[14]It is interesting to notice that even though Jeffreys' arguments are often considered to have played a significant role in the rejection of Wegener's hypothesis, Oreskes suggests that they "proved quite insufficient to move most geologists" (Ibid., p. 89). She mentions both British and North American geologists who were either not inclined to fully accept Jeffereys' views or who did not have too high opinion of him as a scientist. Also see (Oreskes 2001, p. 218).

was needed to confirm the relation between Drift and seismology, and that the conflict between the two was not at all inevitable.

We can thus conclude that Wegener's theory had a programmatic character, which at least temporarily addressed the major problems it was confronted with.

So far we have shown that Drift satisfied all the required standards for being initially WPSS. In the following section we will examine whether it also remained WPSS.

4.8 Theoretical growth and the Growth in the Programmatic Character

In this section we will show that Drift exhibited a theoretical growth throughout the 1920s and the 1930s.

4.8.1 Growth that Drift Exhibited in the 1920s

4.8.1.1 Increase in the Number and Quality of Significant Explanations

Up to the 1920s the main geological evidence, first for Suess' hypothesis of Gondwana and then for Wegener's Drift were the similarities between Karroo formations in South Africa and age-equivallent rocks elsewhere in the world (Oreskes 1999, p. 157). However, a direct comparative study of the so-called Gondwana beds was missing, and both Suess and Wegener built their ideas by combining results obtained by other geologists, rather than themselves conducting a field investigation. North American geologist Frederick Wright realized that the similarities between these regions could thus be taken as a prediction of Drift. He proposed that the examination of the evidence be conducted by an expert in this field, namely Alexander Logie du Toit, a leading specialist in the geology of South Africa who thus had a sufficient expertise for comparing South African coast with the South American one (Ibid., p. 158). The proposal was accepted and in 1923 du Toit embarked on a journey to South America, sponsored by the Carnegie Institution of Washington, to study the geology of the eastern coast of the continent. The results of his study were, according to du Toit himself, strikingly in favor of the Drift hypothesis. Litho-stratigraphic characteristics of the South American east coast were so similar to those of the South African coast that du Toit concluded the two continents must have been at one time no more than 400-800 kilometers apart (Oreskes 1999, p. 161). For example, the facies patterns on both sides of the Atlantic exhibited less change when compared to each other than to much closer facies within their respective continents (Ibid., p. 166). According to du Toit, this evidence required direct physical proximity of the continents, and could thus be explained only by Drift. The study thus greatly contributed to the increase of significant explanations offered by Drift. The results were presented in Du Toit's monograph *A Geolog-*

ical Comparison of South America with South Africa published in Washington, D.C. in 1927.

In addition to du Toit's work, Drift showed a growth in the work of some other geologists as well. Swiss geologist Emile Argand improved Wegener's solutions of orogenic problems by offering a more detailed account of the formation of mountain ranges and island festoons ((Frankel 1981, p. 202); (Oreskes 1999, p. 115)). Arthur Wade, a geologist who first lived in England and then in Australia, approached Drift from the perspective of his oil exploration work. He found Drift to be fruitful for accounting for the structure and history of New Guinea, whose crustal deformation he attributed to its mashing against Australia. Calling upon Drift even had practical consequences in this case, since it could help in locating sites for future oilfields (Le Grand 1988, p. 86).[15] The explanatory power of Drift was also improved by the work of some Australian biologists. For example, zoologist Launcelot Harrison considered the problems of Australian bio-geography and suggested that if the land connections between the southern continents had to be rejected on geophysical grounds, Drift was the only remaining hypothesis offering an explanation of species distribution in the South (Ibid., p. 87).

4.8.1.2 Improved Programmatic Character – The Mechanism of Drift

As we have seen, Wegener addressed the problem of the mechanism of the drift by calling upon the theory of isostasy, and in addition, by hypothesizing two forces that could be responsible for the continental movements. Nevertheless, neither of these forces turned out to be sufficiently strong to account for the drift. Wegener eventually had to distance himself from this question and to admit that "The Newton of drift theory has not yet appeared." (Wegener 1966, p. 167). The opponents argued that finding such a force was improbable, which introduced an external probabilistic inconsistency between Drift and the physical theory. Nevertheless, throughout the 1920s Wegener's followers (Reginald Daly, John Joly and Arthur Holmes) offered possible solutions for this problem. We will take a closer look at the most important of these accounts – Arthur Holmes' model of Drift.

Among all the models of Drift from the first half of the twentieth century, the account offered by British geologist Arthur Holmes in the late 1920s most successfully addressed both the question as to whether Drift was conceivable in light of the physical theory, as well as the question as to which forces could be responsible for such a movement of continents (see (Frankel 1978, p. 131)). Coming from the field of radiology, Holmes argued that due to the thermal processes resulting from the radioactive materials in the inside of the earth, there was an accumulation and a discharge of the heat. Thus, on the one hand,

[15] Wade made further contributions in the 1930's after emigrating to Australia, where he worked on the geology of Western Australia. His research showed that there was a fitting of the southern continents to Antarctica as well as the matching-up of some of their geological features (Ibid., p. 86).

he disqualified the basic idea of the contractionist tradition – the hypothesis of the cooling earth. On the other hand, he proposed the idea of the convection currents in the substratum, resulting from the differential heating by radioactivity. His model has been sometimes labeled as "seafloor thinning": it supposes that the continents drift apart by being carried along the backs of the convection currents, which arise beneath continents, diverge and move towards the continental edge; as a result, the currents produce a "stretched region" of crustal material, which eventually becomes a new ocean floor (Ibid., p. 131-143). In addition to being compatible with the results of the research in radiology, his hypothesis had a strong explanatory power: it could account not only for the features of the continental drift, but also for the phenomena such as mountain building, oceanic deeps, geocynclines, rift valleys, the distribution of earthquakes and volcanos, etc. (Holmes 1931, p. 600).[16] Holmes thus managed to address both, the alleged conflict between Drift and physics (by showing that the continental drift is possible even without assuming that continents plow their way through the seafloor), as well as the question of the forces governing the drift (by proposing the forces of the drift, namely, the convection currents in the earth's mantle).

However, Holmes was careful enough to see his hypothesis as a "preliminary survey":

> So far the treatment has been almost entirely qualitative and therefore it inevitably stands in need of criticism and quantitative revision. The hydrodynamics of the substratum and its behaviour as a heat engine need to be attacked on sound physical lines. The capacity of substratum currents to promote magmatic corrosion, transport and crystallisation, and to produce migrating sub-crustal wave forms, calls for detail treatment. The full bearings of the hypothesis on petrogenesis have yet to be investigated. Meanwhile its general geological success seems to justify its tentative adoption as a working hypothesis of unusual promise. (Ibid., p. 600)

We see here a number of tasks that Holmes points out as relevant for further examination of his hypothesis. Hence, the idea of the convective currents in the substratum can be taken as a prediction of his model of Drift, which could be investigated in different ways in future field work.

To sum up, Holmes' model obviously gave a programmatic character to the theory with respect to the problem of the mechanism of Drift. Even Harold Jeffreys, one of the biggest opponents of Drift, had to admit that Holmes' proposal rendered the idea of Drift "physically possible" (though he still found it to be very implausible) (Frankel 1978, p. 147).

[16] Note that these features of Holmes' model (its compatibility with radiology and a strong explanatory power) are not directly relevant for the programmatic character of Drift, though they will turn out to be important for the criterion of external inferential density.

4.8.1.3 Improved Programmatic Character – Seismology

In addition to Wegener's replies to Jeffrey's objections (see Section 4.7.2), other proponents of Drift also discussed this issue. Already in the early 1920s Reginald Daly argued that this objection to Drift conflated rigidity with solidity (Oreskes 1999, p. 93). The results of seismic studies supported the idea that the substrate was rigid, but this did not imply that it was solid (similar to the properties of glass, which is solid at room temperature, but which under pressure and over time actually flows). Daly further explicated that when it comes to properties of the substrate which are relevant for Drift, what mattered was not the distinction between liquidity and solidity, but the one between crystalline and non-crystalline materials. And if the substrate were non-crystalline (like glass), it could appear as rigid in response to seismic waves, but plastic in response to long-term effects (Ibid., p. 93-94). John Joly had a different response to this objection: according to his model, the substrate beneath the continents was periodically and locally (rather than continuously and uniformly) molten (Ibid., p. 108).

Therefore, the proponents of Drift offered further possible solutions for the conflict between their theory and seismology, the validity of which depended on further investigations of the properties of the substrate.

4.8.1.4 Increase in the Internal and External Inferential Density

As we have seen, Holmes' model of Drift provided a more unified core of hypotheses than Wegener's one in view of the growing criticism of the latter for the lack of plausible forces of drifting.[17] Furthermore, through Holmes' model (and previously through Joly's model as well) Drift obtained inferential connections with the theory of radioactivity, which became problematic for both of its rivals. The incompatibility of radioactivity with either contractionism or permanentism was the primary reason for Holmes to become a proponent of Drift, since it was the only theory that could account for the accumulation and discharge of heat, necessitated by the presence of radioactive materials.

Drift showed an additional growth by improving its inferential connections with the theory of isostasy. In the second half of the 1920s William Bowie, a proponent of permanentism, organized an international collaboration aimed at investigating isostasy, but indirectly relevant for the hypothesis of drift as well (Oreskes 1999, p. 236-261). The aim of the investigation was to test the theory of isostasy by obtaining gravity data from the ocean floor. Up to that point isostasy had been confirmed only on the continents due to the fact that there were no precise instruments for measuring gravity at sea. In the meantime, Dutch geologist Felix Vening Meinesz developed an improved gravimeter, suitable for the sea measurements as well. The results of the investigation conflicted with the assumptions of Pratt's model of isostasy, used in the permanentist conception of geology. Indirectly, it indicated that Airy's

[17] For a detailed discussion of the unifying aspect of Holmes model see (Lewis 2002).

model of isostasy, compatible with Drift, might be correct after all. Thus, the inferential connections between Drift and isostasy were in this way improved.

4.8.2 Growth that Drift exhibited in the 1930s

4.8.2.1 Increase in the External Inferential Density

Expeditions organized in the 1930's corroborated the above mentioned results of the research on isostasy, which were in accordance with Airy's model, and thus with Drift as well (Oreskes 1999, p. 245). By the mid-1930s seismic evidence refuted Pratt's model of isostasy (Ibid., p. 258). In view of these results, American geologist Richard M. Field pointed out in 1937 that Wegener's hypothesis played a great role in stimulating geological and geophysical investigations (Ibid., p. 259). Moreover, Bowie himself acknowledged in 1936 the link between these results and Wegener's ideas:

> The Wegener hypothesis has received a great deal of attention in recent years and deservedly so. It is based upon the idea [of] isostasy .... Many students of the Earth's crust feel that the Wegener hypothesis does violence to certain mechanical principles, but, in any event, it is something that should be looked into. (quoted from Ibid., p. 261).

4.8.2.2 Further Improvement of the Programmatic Character - The Mechanism of Drift

Vening Meinesz in a volume from 1934, which discussed results of his investigations of gravity, addressed some of Jeffreys' objections to the hypothesis of convection currents as the force governing the drift. According to Jeffreys, any significant thermal differentials in the earth must have been eradicated throughout its cooling history. However, Meinesz pointed out that the heterogeneous nature of the earth's inner structure, with an uneven distribution of radioactive constituents and thermal properties meant that "in the actual earth there can be no doubt that convection currents must develop" (quoted from Ibid., p. 245-248). Meinesz's work offered support to the hypothesis of Holmes' model, which, as we have seen, gave a programmatic character to Drift with respect to the problem of the mechanism of drifting.

4.8.2.3 Increase in the Quality of Significant Explanations

Some explanatory anomalies pointed out by the opponents of Drift were addressed by Alexander du Toit's capital work *Our Wondering Continents*, which came out in 1937 (du Toit 1937). For example, in order to account for more explananda, Du Toit proposed in place of Wegener's one super-continent, two original super-continents – Laurentia in the north and Gondwanaland in the south.

4.8.3 State of affairs in the 1940s

In spite of exhibiting a theoretical growth and a growth of its programmatic character throughout the 1920s and the 1930s, Drift was rarely discussed in the '40s ((Le Grand 1988, p. 117); (Oreskes 1999, p. 226)). This was partially due to the effects of World War II, during which many North American geologists were employed in the war effort, while the immediate post-war years were not very fruitful of theoretical developments.

The development of Drift seemed to have reversed by the articles published by North American paleontologist George Gaylord Simpson. However, Simpson's objections were rebutted by du Toit ((Oreskes 1999, p. 295-296); (Frankel 1987, p. 217-219)).[18]

To sum up, we have shown that Drift had a theoretical growth throughout the 1920s and 1930s. In spite of this growth, it received hardly any attention in the 1940s. It is thus not surprising that its theoretical growth was missing in this decade.[19]

4.9 The Consequences of Our Account for the Epistemic Stances of Geologists

Our analysis has shown that Drift scored well in each of the criteria of our evaluation. Thus, we can conclude that it was both initially WPSS and that it remained WPSS in the first half of the twentieth century. That means that it was rational to consider Drift as worthy of pursuit, and that it was irrational to reject its pursuit as unworthy. It also means that characterizing Drift as worthy of pursuit was not conflicting with rejecting its full acceptance.

In this section we are going to show that most of those who had a positive opinion of Drift actually found it to be worthy of pursuit. Furthermore, we will also show that there were geologists whose opinions of Drift can be considered irrational in the above explicated sense.

4.9.1 The Supporters of Drift

Even though Wegener himself had a strong epistemic stance towards his own theory, most of those who argued in favor of Drift, maintained that it was a

[18] Moreover, in the mid-40s du Toit began a work on a manuscript entitled "On the mathematical probability of continental drift" (Ibid., p. 297) in which he planned to offer a quantitative account of Drift based on the degree of similarity among species and the distances among them. The work was never finished due to du Toit's death in 1948. Thus, this work cannot be considered as a contribution to the explanatory growth of Drift, but it does represent a contribution to its programmatic character in the time when Drift was mainly abandoned.

[19] Note that this conclusion differs from R. Laudan's view according to which "far from showing a greater rate of progress than rival theories, drift stood still, or even regressed between 1930 and 1955" (Laudan 1987, p. 214). In view of our analysis, Laudan's estimation is too rough since it does not apply to the 1930s.

theory requiring and worthy of more investigation. Let us mention some of them.

In Europe, German paleontologist Karl Andree found Wegener's theory to be a stimulus to research even though it could not be accepted in all of its details (Le Grand 1988, p. 58). Austrian paleontologist Bruno Kubart maintained that a combination of ideas taken from older theories together with those from Wegener's Drift could form a suitable basis for further research (Ibid., p. 60). Dutch geologist Gustaff A. F. Molengraff argued that eastward drift was a possibility (Stewart 1990, p. 37). For a British geologist Charles Seymour Wright, Drift offered a promise since it could explain certain fossil deposits on Antarctica which indicated that previous to glaciation there was a period of warmth in this area (Le Grand 1988, p. 89-90). Irish geologist John Joly suggested that Drift was a logical possibility within his theory of periodic convection currents (Stewart 1990, p. 37). North American geologist Chester Longwell pointed out that "if the doctrine of continental displacement is accepted as a working hypothesis, to be tested and tried fairly along with others, it may be productive of valuable results" (quoted from (Stewart 1990, p. 38)). Joseph T. Singewald suggested that, in spite of the obvious failures of Wegener's presentation, the hypothesis should be tested on the basis of its worth for guiding research (Le Grand 1988, p. 71). Even Arthur Holmes found Drift to be a possible working hypothesis rather than a theory sufficiently developed to be accepted (Stewart 1990, p. 41).[20] Leo Arthur Cotton, Australian geologist, found Drift (though not Wegener's version) to be worthy of pursuit or entertainment with respect to the problems that concerned him (Le Grand 1988, p. 85). Arthur Wade, who was educated in England, and afterward emigrated to Australia and who was engaged in oil exploration around the world, characterized Drift as a working hypothesis and pointed out its application to economic geology of those regions in which he had conducted his research (Ibid., 86). Similar was the opinion of Australian zoologist Launcelot Harrison who found Drift explanatory of the southern species distribution and thus to be a useful working hypothesis (Ibid., p. 88).

As we have mentioned, the positive stance towards pursuit of Drift is not necessarily conflicting with rejecting Drift in the context of acceptance. It is easy then to see that debates among the proponents of Drift advocating its pursuit and the opponents rejecting its full acceptance sometimes consisted of not necessarily conflicting arguments. That means that the awareness of the distinction between theory evaluation in the context of pursuit and theory evaluation in the context of acceptance may sometimes help scientists to avoid unnecessary debates. In other words, the question of pursuit worthiness is not

[20]Stewart remarks that the fact Drift was for Holmes only a possible working hypothesis shows that Holmes's stance towards it was very weak and can hardly be seen as the one of a strong supporter of the theory, who would encourage his colleagues and students to advocate a novel and widely opposed theory (Ibid., p. 42). However, Stewart's conclusion shows that he does not recognize that judging a theory as worthy of pursuit represents a valuable contribution to its further development, even though such a stance may very well be the most rational form of supporting a newly developing theory.

only of significance for philosophical discussions regarding issues of rationality, but it is also of significance for scientific practice and epistemic stances of scientists.

Among the opponents of Drift there were also those who rejected not only its acceptance, but also its pursuit worthiness. Let us take a closer look at their points of view.

4.9.2 Opponents who Rejected Pursuit Worthiness of Drift

That Drift was not always acknowledged as worthy of pursuit is exemplified in the opinions of geologists who explicitly ridiculed it. For example, Bailey Willis' 1944 article was titled "Continental Drift, Ein Märchen" – a fairytale. As Le Grand puts it: "His hostility to Drift, even as a permissible working hypothesis for other geologists, was unabated in 1944" (Le Grand 1988, p. 118). Similarly, Charles Schuchert still spoke of "the Wegener sliding circus" in 1931 (Oreskes 1999, p. 212), while Max Semper explicitly rejected the idea of pursuing this "absurd theory" (Le Grand 1988, p. 59). Even in the 1950s advocates of Drift were still publicly ridiculed (Oreskes 1999, p. 218). These opponents not only found Drift to be unworthy of pursuit, but disregarded it even as a serious scientific theory.[21] A similar attitude towards Drift can be found also in later discussions. For example, geophysicist Seiya Uyeda suggested that Drift could scarcely be regarded as scientific since it could not explain what had originally caused the continental movements (Oreskes 1999, p. 63, fn. 28).[22] All of these views strongly diverge from the result of our analysis.

But we have to pause here and take into consideration a possible objection that geology in the first half of the twentieth century had different methodological standards, and that thus our criteria of pursuit worthiness are not applicable to the notion of rationality governing scientific research at the time. More precisely, North American geology in the first half of the twentieth century was rooted in a methodological framework which was deeply embedded in inductivist ideals. Many authors who discussed this historical episode suggested that these geologists primarily focused on field research and practically valuable results, placing less significance on global geological theories and their explanatory power. And if this is correct, then their criteria for what counts as epistemically worthy of pursuit might have been different as well.

However, Naomi Oreskes shows that the view according to which North American geology was deeply inductivist and anti-theoretically driven is in fact a historiographic cliche, and that describing these geologists as naive empiricists or narrow utilitarians doesn't do justice to their research. Not only

[21] As Frankel remarks: "I do not find it surprising that they would not accept the drift hypothesis, but I do find it surprising that they would not treat it as a serious research program." (Frankel 1976, p. 319).

[22] Note that Uyeda's epistemological standard, requiring for a deepening of explanations offered by a given theory as the condition for it to be regarded as scientific, differs from our standards which allow for problems of this kind to be tackled by the programmatic character of the pursued theory. For further discussion on the validity of Uyeda's standard see: Ibid,, p. 63-64.

were they not opposed to theoretical activities as such, but some of the major theoretical contributions to earth sciences came from the United States (for example, James Dana's work on the origin of continents and oceans, James Hall's geosyncline theory, or Clarence Dutton's work on isostasy) (Ibid., p. 129). Furthermore, Thomas C. Chamberlin, one of the most important American geologists from this time period, promoted the unity of theory and practice (Ibid., p. 130-133).

Where American geologists differed from European ones was in their suspicion of theory-driven science and in requiring a thorough empirical research as a necessary step preceding any theoretical claims (Ibid., p. 134-136). A direct observational statement of geological phenomena was to come before any theoretical conclusions. Moreover, the research was to be done as much as possible in terms of G. K. Gilbert's and T. C. Chamberlin's method of multiple working hypotheses. As the name suggests, the underlying idea of the method was to view observational facts in light of competing explanatory frameworks, rather than in view of an already established theory. The goal of the method of multiple working hypotheses was to navigate between the risks of dogmatic deductivism and the infertility of naive inductivism (Ibid., p. 140).

In view of such methodological standards, Oreskes argues that the key reason why North American geologists reacted so negatively to Wegener's theory is the fact that Wegener violated these standards in several respects. First, his program aimed at proposing a grand geological theory. Second, he regarded the supporting evidence as "proofs" necessitating Drift, rather than as observations or geological facts which were best explained by his theory. Finally, he presented the idea of Drift not as a working hypothesis, but as a "fundamentally correct" theory, in contrast to contractionism and permanentism which he saw as based on erroneous premises (Ibid., p. 153-154). As a result, some American geologists not only rejected Drift, but found Wegener's approach to be unscientific.[23]

It is not difficult to understand then why Wegener's approach was not appealing for North American geologists. The fact that his theory violated the standards of how science is to be done in their view explains why his theory could not be accepted at the time. But was Drift, in view of these standards, also unworthy of pursuit? All that the above mentioned objections show is that Wegener might have been incautious and that he might have had an unjustified epistemic stance towards his own theory. But they do not attack the fact that Wegener's theory exhibited an explanatory power for a certain set of phenomena. The closest Wegener's opponents came in criticizing the fact that Wegener's theory offered some explanations was to argue that he "generalized too easily from other generalizations" (Stewart 1990, p. 37). Nevertheless, they did not mind that their own "generalizations" depended on ad hoc hypotheses – for example, on the idea of land bridges for which they had no mechanism

[23]For example, American geologist Rollin. T. Chamberlin questioned the scientific status of the entire field of geology in view of the fact that it allowed for theories like Drift "to run wild" (Le Grand 1988, p. 64).

which would explain their disappearance (Ibid.). In addition, the explananda addressed by Drift were not merely posited by Wegener, without any empirical back-up. For example, paleontological similarities between coastal regions on the opposite sides of the oceans were researched by others as well, and even aknowledged by Wegener's opponents.[24] Finally, the research conducted by du Toit, Holmes and others introduced much more substantiated evidence and thus improved Wegener's theory in view of American methodological requirements. Therefore, there was no methodological reason why Drift would not be taken seriously as any other working hypotheses. Yet, as we have seen, some geologists found Drift to be unworthy of pursuit even in the 1930s – long after Wegener's original proposal had been significantly improved. The root of their epistemic stance can thus be found primarily in their biasness towards the fixist frameworks, rather than in a fair application of a specific set of methodological standards.[25] As a matter of fact, our criteria of pursuit worthiness do not conflict in any significant way with these methodological standards. In contrary, they are compatible with the underlying idea of the method of multiple working hypotheses: they allow for a simultaneous pursuit of different hypotheses, since more than one theory (or a research tradition) can be, according to our framework, WPSS.

We can thus conclude that, in spite of the methodological differences among geologists, the opinion that Drift was worthy of pursuit in the first half of the twentieth century, and especially in the 1920s and the 1930s, can be characterized as rational (in the sense of being epistemically justified), and the rejection of its pursuit worthiness as irrational.

4.10 Conclusion

In this chapter we have presented en epistemic evaluation of the pursuit worthiness of Drift in its early development. For this purpose we have used the framework of epistemic justification suitable for theory evaluation in the context of pursuit, which we have adapted for the evaluation of pursuit worthiness

[24] For instance, Charles Schuchert, very critical of Wegener's theory, acknowledged at the 1926 American Association of Petroleum Geologists symposium on continental drift "that Wegener's hypothesis has its greatest support in the well known geologic similarities on the two sides of the Atlantic, as shown in strikes and times of mountain-making, in formational and faunal sequences, and in petrography." (quoted from (Oreskes 1999, p. 180)). Ironically, Wegener was actually attacked for using results of the research conducted by others, instead of doing all the field work on his own since, as Schuchert remarked, "it is wrong for a stranger to the facts he handles to generalize from them to other generalizations" (quoted from (Oreskes 1999, p. 156)).

[25] A biased approach of North American geologists is also reflected in the fact that their judgment was made in view of locally relevant sets of explananda, disregarding geological phenomena belonging to other regions around the world, for which Drift was highly explanatory. Le Grand calls such an approach "localism" (Le Grand 1988, p. 95-97), while Oreskes characterizes it as "epistemological chauvinism" or "epistemological affinity"(Oreskes 1999, p. 52-53), pointing out that placing a higher preference on certain subsets of the available data was often motivated not only by a specific geographical context, but also by a national or disciplinary one.

in a strong sense. We have shown that Drift had a number of significant explanations, that it did not have a lower internal or external inferential density than its rivals, and that it had a programmatic character with respect to its major problems. Moreover, we have shown that throughout the 1920s-'30s Drift exhibited a theoretical growth and a growth in its programmatic character, and thus remained worthy of pursuit in a strong sense throughout this time period. On the one hand, this means that it was not only rational to pursue Drift, but that characterizing Drift as worthy of pursuit was not conflicting with rejecting its full acceptance. Hence, we have emphasized that the distinction between theory evaluation in the context of acceptance and the one in the context of pursuit may help scientists to avoid some unnecessary debates. On the other hand, we have shown that it was epistemically unjustified to reject Drift as unworthy of pursuit, and that consequently, opinions of some geologists in the first half of the twentieth century can be regarded as irrational.

It is important to clarify that our analysis did not take into account the question of pursuit worthiness of other rivaling theories at the time. However, it may very well be the case that a closer look at contractionism and permanentism would reveal that they were also worthy of pursuit. A detailed evaluation of each of them remains a task for the future research.

Chapter

5 Argumentative Shift in Methodology

✎ *I wish to thank Christian Straßer for the valuable comments on a former version of this chapter.*

Summary In this chapter we will address some questions concerning possible relativism underlying the evaluation of epistemic pursuit worthiness. Since the problem of relativism in theory evaluation is usually connected to Kuhn's views, we will take a closer look at his stance on theory evaluation in the context of pursuit. We will show that the idea of persuasion and argumentative reasoning, proposed by Kuhn, can help in reducing possible relativistic outcomes. We will then track these two ideas in works of Mcmullin and Pera: while, McMulllin extends Kuhn's notion of persuasion from the level of scientific theories to the meta-theoretic level, that is, the level of methodological standards, Pera uses Kuhnian persuasion as a motivation for an argumentative approach to methodology in general. By showing the significance of argumentative reasoning in theory evaluation, we will motivate a specific argumentation-based approach to formal modeling of scientific reasoning.

5.1 Introduction

The previous two sections presented and exemplified the epistemic evaluation of the pursuit worthiness of theories regarding the rationality underlying a given scientific community. We finished the discussion on Drift by examining whether the difference in epistemic standards used by the proponents of the rivaling research traditions could pose a problem for our evaluation. To this end, we took a closer look at the methodological discussions at the time, as well as at the arguments constituting the debates between the scientists. In this way we could determine which types of reasons were used to attack Drift, and whether these reasons indicated different epistemic standards of theory evalu-

ation or not. This discussion points to two more general questions regarding the evaluation of the epistemic communal pursuit worthiness. First, we may ask which possible obstacles can stand in the way of a non-relativistic outcome of such an evaluation, and if there are ways in in which such obstacles could be surpassed. Second, we have so far presented this type of evaluation only in an informal way. But what about the formal modeling? What kind of formal modeling would be suitable for this task?

The aim of this chapter is to shed some light on these two questions. By answering the former question (or at least, pointing to the direction in which it could be tackled), we will also show important properties that the formal modeling of theory evaluation in the context of pursuit should allow for. With regards to the first question, the problem of a relativistic theory evaluation immediately points to a Kuhnian perspective. Hence, in the following section we will take a look at the status of theory evaluation in the context of pursuit in Kuhn's writings. Kuhn's concept of persuasion will turn out to be an important tool in facilitating the debates regarding pursuit worthiness of rivaling paradigms. In subsequent sections we will show that the role of argumentative reasoning in theory evaluation has been also pointed out by McMullin and Pera. In view of this discussion we will suggest that if formal modeling of theory assessment, suitable for the context of pursuit and the context of acceptance, should replace the old ideal of a rule-based approach, and instead reflect the dynamics of argumentative reasoning, it has to allow for an assessment in view of arguments constituting scientific debates.

5.2 Kuhn and the Context of Pursuit

The main obstacle in understanding Kuhn's stance on pursuit worthiness is his neglect of the distinction between the context of pursuit and the context of acceptance. Both contexts are discussed under the topic of theory choice, that sometimes regards the early stages of a development of a paradigm, and sometimes the later stages in which one paradigm is replaced by another. Despite the lack of an explicit distinction between them, there are places in which Kuhn clearly addresses the question of theory pursuit. In this section we will try to explicate his stance on this issue by closely reading his writings (especially the 1969 *Postscript* to *The Structure*[1]). I will argue that Kuhn's views on theory change and rationality that underlies it do not render the evaluation of pursuit worthiness as relativistic as the evaluation of theories in the context of acceptance. To this end, I will show:

1. first, that Kuhn often describes theory evaluation in the context of pursuit in terms that are more suitable for theory evaluation in the context of acceptance;

[1] Published in (Kuhn 1996, 174-210).

2. second, that he makes an implicit distinction between the idea of the initial and the successive pursuit worthiness (which has been explicated in Chapter 3;

3. third, that he does not clearly distinguish between group and individual rationality in the context of pursuit and that his discussion of theory choice in the context of pursuit primarily regards the notion of individual pursuit worthiness without being distinguished as such;

4. fourth, that Kuhn provides certain tools that may be helpful in achieving an agreement regarding the communal pursuit worthiness, but reserves them only for the evaluation of the successive pursuit worthiness.

Let us begin with Kuhn's notes on the context of pursuit from *The Structure*:

> ... if a new candidate for paradigm had to be judged from the start by hard-headed people who examined only relative problem-solving ability, the sciences would experience very few major revolutions ... But paradigm debates are not really about relative problem-solving ability, though for good reasons they are usually couched in those terms. Instead, the issue is which paradigm should *in the future* guide research on problems many of which neither competitor can yet claim to resolve completely. A decisions between alternate ways of practicing science is called for, and in the circumstances that decision must be based less on past achievement than on *future promise*. The man who embraces a new paradigm at an early stage must often do so in defiance of the evidence provided by problem-solving. He must, that is, *have faith* that the new paradigm will succeed with the many large problems that confront it, knowing only that the older paradigm has failed with a few. *A decision of that kind can only be made on faith.* (Kuhn 1996, p. 157-158, italics added)

The parts in italics indicate that Kuhn here obviously speaks of theory evaluation in the context of pursuit. Discussions regarding pursuit worthiness of a new candidate for a paradigm concern the future promise of a theory. Such a promise is based on faith that the paradigm will succeed in solving its current problems. Kuhn goes on to argue that a crisis is important precisely in order to allow for a new candidate to be at all noticed. He then adds:

> But crisis alone is not enough. There must also be *a basis, though it need be neither rational nor ultimately correct*, for faith in the particular candidate chosen. Something must make *at least a few scientists* feel that the new proposal is on the right track, and sometimes it is *only personal and inarticulate aesthetic considerations* that can do that. Men have been *converted* by them at times

> when most of the articulable technical arguments pointed the other way.
>
> This is not to suggest that new paradigms triumph ultimately through some mystical aesthetic. On the contrary, very few men desert a tradition for these reasons alone. Often those who do turn out to have been misled. But if a paradigm is ever to triumph it must gain some first supporters, men who will develop it to the point where hardheaded arguments can be produced and multiplied. And even those arguments, when they come, are not individually decisive. Because scientists are reasonable men, one or another *argument* will ultimately *persuade* many of them. But there is no single argument that can or should persuade them all. (p. 158, italics added)

There are three points that should be noted in view of the above quote:

1. First, the quote indicates that Kuhn subscribes properties that are more characteristic for the context of acceptance to the evaluation in the context of pursuit. On the one hand, immediately after mentioning that there needs to be a certain basis for the faith in a new paradigm, he adds that the reasons forming such a basis have converted scientists, in spite of the counterarguments. By linking the reasons used for arguing that a theory is worthy of pursuit with the epistemic stance of convergence[2] which amounts to a full theory acceptance, Kuhn implicitly overlooks a more moderate stance towards pursuit worthiness. For, why would he otherwise go as far as to speak of conversion, in order to show that, in spite of criticism, a young theory may still be considered worthy of pursuit? As we have pointed out many times, a scientist may assess a paradigm as worthy of pursuit (for the given scientific community) without engaging in its actual pursuit. On the other hand, Kuhn writes that the basis for pursuit worthiness which he explicates in this quote should not suggest "that new paradigms triumph ultimately through some mystical aesthetic". Once again he directly links the context of pursuit with the epistemic stance that is typical for context of acceptance: a "triumph" of a paradigm indicates its superiority towards other paradigms, which is not at all needed for it to be considered as worthy of pursuit.

2. Second, Kuhn here seems to distinguish between two stages of the evaluation of pursuit worthiness of a paradigm. On the one hand, *the first supporters* of a paradigm *feel* that the new proposal is on the right track. Kuhn here departs from rational reasons one might have for evaluating a new paradigm as initially worthy of pursuit. He allows for "personal and inarticulate aesthetic considerations", which may turn out to have

[2] Kuhn distinguishes *persuasion* – a process in which one person convincing another that the former one's view is superior and ought therefore supplant the view of the latter (p. 203) from *conversion* – a process in which one is not only persuaded, but also internalizes the new theory and is "at home in the world it helps to shape" (p. 204).

been misleading, to serve as the basis of the evaluation. On the other hand, if the first supporters develop the paradigm further, it may reach the point where arguments can finally be used. This is the second stage of the evaluation of pursuit worthiness, which proceeds in rational terms, and where the process of persuasion can begin.

3. Third, when speaking of pursuit worthiness, Kuhn seems to focus exclusively on the notion regarding the individual scientists, rather than on the notion regarding the given scientific community.[3] He claims that on the basis of certain reasons *some* scientists may find a new theory worthy of pursuit. Moreover, by linking the evaluation of pursuit worthiness with the notions indicating theory acceptance (conversion, triumph), he seems to assume that if a scientist evaluates a theory as worthy of pursuit, she will also engage in its actual pursuit, since she has found it superior to other candidates. Such an approach refers to the "individual" pursuit worthiness, rather than to the "communal" one.[4]

We will now take a closer look at the latter two points.

5.2.1 The Criteria of Pursuit Worthiness

We begin with the second point, regarding different stages of the evaluation in the context of pursuit. In response to the criticism accusing him of irrationality underlying theory choice, Kuhn argues in the *Postscript* that his stance does not entail that proponents of incommensurable paradigms cannot communicate with each other, nor that in their debates there can be no recourse to good reasons (Kuhn 1996, p. 199). He rejects accusations according to which such reasons are ultimately personal, subjective or irrational. He explains that his point was rather to show that debates over theory choice cannot be cast in a form that fully resembles a logical or mathematical proof (p. 199).

If these comments are also applicable to the context of pursuit (and as his remarks in (Kuhn 1977, 320-339) show, they are), Kuhn then seems to argue that his treatment of the evaluation of pursuit worthiness of paradigms is not based on reasons that are *ultimately* personal or subjective. Indeed, once the first supporters of the paradigm have developed it, the process of persuasion can begin. What Kuhn rejects is that this subsequent stage of pursuit worthiness evaluation can be given in a form or a logical or mathematical proof. Hence, he address the latter of the two stages in the evaluation of pursuit worthiness – the successive pursuit worthiness. Let us see his characterization of this stage.

Regarding the evaluation of the successive pursuit worthiness of paradigms, Kuhn rejects what we have elsewhere called *the strong notion of rationality*[5]

[3] For the distinction between the two see Chapter 2.

[4] This is not to say that in order for a theory to be worthy of pursuit for an individual scientist, she needs to go through a process of conversion, or even to assess the young paradigm as superior to its rivals. These are just some of the sufficient (but not necessary) reasons for an individual pursuit worthiness.

[5] See (Šešelja and Straßer 2009, p. 323), also printed in the Appendix of this thesis.

governing theory choice. According to this notion, by applying the criteria that are shared by the scientific community, one obtains a unique choice of a scientific theory. In other words, this notion assumes that all scientists will have the same preference order on theories on the basis of a shared set of criteria with which they evaluate them. Such an idea of rationality Kuhn compares to an "algorithmic path" or "a logical or mathematical proof" (Kuhn 1996, 198-199). According to him, discussions among scientists instead of such an algorithmic path often take the path of *persuasion*:

> ...the superiority of one theory to another is something that cannot be proved in the debate. Instead, I have insisted, each party must try, by persuasion, to convert the other. ...Nothing about that relatively familiar thesis implies either that there are no good reasons for being persuaded or that those reasons are not ultimately decisive for the group. Nor does it even imply that the reasons for choice are different from those usually listed by philosophers of science: accuracy, simplicity, fruitfulness, and the like. What it should suggest, however, is that such reasons function as values and that they can thus be differently applied, individually and collectively, by men who concur in honoring them. (p. 198-199)

Kuhn here tells us that the process of persuasion is formulated in terms of arguments that call upon a set of epistemic values. In his (Kuhn 1977) he specifies these values as a set of shared criteria or *the* shared basis for theory choice:

> "These five characteristics – accuracy, consistency, scope, simplicity, and fruitfulness – are all standard criteria for evaluating the adequacy of a theory. ...I agree entirely with the traditional view that they play a vital role when scientists must choose between an established theory and an upstart competitor. Together with others of much the same sort, they provide *the* shared basis for theory choice." (Kuhn 1977, p. 322)[6]

Taking into account Kuhn's remarks on the context of pursuit that we have previously mentioned, we can now say that persuasion is a process in which scientists try to convince one another, by means of shared epistemic values, that a certain theory is promising of being further investigated. And precisely the fact that criteria of theory choice function as values, rather than rules, allows them to function as criteria also "in the earliest stages of theory choice" (Kuhn 1977, p. 321), that is, in the evaluation of the successive pursuit worthiness of theories.

However, Kuhn points out that "There is no neutral algorithm for theory-choice, no systematic decision procedure which, properly applied, must lead

[6]See also (Kuhn 2000, p. 96).

each individual in the group to the same decision" (Kuhn 2000, p. 200, italics added). This is due to the fact that "Individually the criteria are imprecise: individuals may legitimately differ about their application to concrete cases."(Kuhn 1977, p. 322). Furthermore, even though the pursuit worthiness is evaluated in terms of a set of shared values, in order to explain why an individual scientist is an early convert to a new system, we need to look at idiosyncrasy as well (Ibid.).[7]

Let us summarize Kuhn's point in view of these quotes. According to Kuhn, scientists initially begin to pursue a paradigm due to, so to say, a gut feeling, inarticulate aesthetic considerations or personal reasons. Once a paradigm has shown a number of results, its pursuit worthiness can be evaluated on the basis of shared epistemic standards. However, different scientists may find different theories worthy of pursuit, due to the fact that they may apply the shared criteria differently, and thus have different preference orders on given theories. This brings us to our third point: the pursuit worthiness regarding individual scientists.

5.2.2 Individual and Communal Pursuit Worthiness

As the above quotes already suggested, when it comes to the context of pursuit, Kuhn's primary focus is set on the pursuit worthiness regarding the individual rationality. Moreover:

> Before the group accepts it, a new theory has been tested over time by the research of a number of men, some working within it, others within its traditional rival. Such a mode of development, however, *requires* a decision process which permits rational men to disagree, and such disagreement would be barred by the shared algorithm which philosophers generally have sought. If it were at hand, all conforming scientists would make the same decision at the same time. With standards of acceptance set too low, they would move from one attractive global viewpoint to another, never giving traditional theory an opportunity to supply equivalent attractions. With standards set higher, no one satisfying the criterion of rationality would be inclined to try out the new theory, to articulate it in ways which showed its fruitfulness or displayed its accuracy and scope. I doubt that science would survive the change. (Kuhn 1977, p. 332)

In view of the individual pursuit worthiness, different scientists may find different theories more worthy of pursuit than others, and hence each of them may decide to engage in a pursuit of a different theory. In order to allow for such a diversity of pursued theories, Kuhn argues that we need to allow for a disagreement among scientists in their evaluation of pursuit worthiness of the

[7]Note that Kuhn here again uses the terminology characteristic for the context of acceptance to describe the context of pursuit: "...idiosyncrasy must be invoked to explain why Kepler and Galileo were early *converts* to Copernicus's system" (Ibid., italics added).

given theories. What Kuhn here completely overlooks is the notion of communal pursuit worthiness. Similarly to others who have conflated these two notions (discussed in Chapter 2), he overlooks the fact that more than one theory may be simultaneously evaluated as worthy of pursuit. On the basis of a shared algorithm for the evaluation of communal pursuit worthiness, not all conforming scientists would have to make the same decision regarding which theory each of them should pursue. They would not have to move, as Kuhn maintains, from one attractive viewpoint to another, never giving a chance to the traditional theory. In contrary, both new candidates as well as the traditional rival could simultaneously be worthy of pursuit for the given community at the time. Hence, by emphasizing a disagreement as a key condition for theoretical diversity in the context of pursuit, Kuhn shows that he not only focuses primarily on the individual notion of pursuit worthiness, but that he also neglects the diversity that can be obtained by means of the communal notion.

But what would be Kuhn's stance on the communal notion of pursuit worthiness? Would he also argue that a disagreement about which theories are worthy of pursuit for the given scientific community is necessary for scientific development, or that it is unavoidable taking into account the specific ways in which the shared criteria are applied by individual evaluators? The following quote sheds some light on these questions:

> ... despite the incompleteness of their communication, proponents of different theories can exhibit to each other, not always easily, the concrete technical results achievable by those who practice within each theory. Little or no translation is required to apply at least some value criteria to those results. (Accuracy and fruitfulness are most immediately applicable, perhaps followed by scope. Consistency and simplicity are far more problematic.) (Kuhn 1977, p. 339)

Here Kuhn seems to suggest that proponents of rivaling theories can, at least partially, evaluate each other's theories in terms of the shared epistemic standards. Moreover, there is a specific "tool" they can use in order to improve their mutual understanding, namely, the process of *translation*:

> However incomprehensible the new theory may be to the proponents of tradition, the exhibit of impressive concrete results will persuade at least a few of them that they must discover how such results are achieved. For that purpose they must learn to translate... (Ibid.)

A more detailed explication of the process of translation can be found in the *Postscript*. With regard to the conceptual differences[8] between the op-

[8]Conceptual differences refer here to Kuhnian incommensurabilities that arise from the differences in similarity relations, which govern the grouping of objects and situations into similarity sets. These sets are "primitive in the sense that grouping is done without an

posing scientific views which cause problems in communication between their respective proponents, Kuhn writes:

> The men who experience such communication breakdowns must, however, have some recourse. ... both their everyday and most their scientific world and language are shared. ... what the participants in a communication breakdown can do is recognize each other as members of different language communities and then become translators. (Kuhn 1996, p. 201-202)

Kuhn explains that by discovering the differences in their respective discourses, the proponents of rivaling paradigms can try to present them to each other by resorting to their shared everyday vocabularies: "Each may, that is, try to discover what the other would see and say when presented with a stimulus to which his own verbal response would be different." (p. 202). The process of persuasion, enriched by the process of translation, thus allows the scientists to see some merits and defects of each other's perspectives.[9]

However, both processes of translation and persuasion are for Kuhn possible only in later stages of the context of pursuit:

> But each language community can usually produce from the start a few concrete research results that, though *describable* in sentences understood in the same way by both groups, cannot yet be *accounted for* by the other community in its own terms. *If the new viewpoint endures for a time and continues to be fruitful*, the research results verbalizable in this way are likely to grow in number. For some men such results alone will be decisive. They can say: I don't know how the proponents of the new view succeed, but I must learn, whatever they are doing, it is clearly right. (p. 203, italics added)

Kuhn here suggests that the initial results of a theory can only be *described* by the scientists working in a rivaling paradigm, but they cannot be *accounted for*, explained in their own terminology. Only if the theory exhibits a growth, some scientists will realize that they should learn the new theory, that is, *step into the circle* of the new paradigm and *learn to speak its language*. Hence, the process of translation and persuasion can begin only with regard to the evaluation of the successive pursuit worthiness.

Nevertheless, Kuhn does not explain why these tools, capable of facilitating rational theory evaluation in the context of pursuit, are applicable only in the later stages of pursuit. If scientists are able to learn the new theory once

---

answer to the question, 'Similar with respect to what?'" (Kuhn 1996, p. 200). In the times of revolutions, some of the similarity relations change. For example, sun, moon, Mars and earth stood in different similarity relations before and after Copernicus (Ibid.).

[9] It is interesting that when Laudan in his (Laudan 1984) criticizes Kuhn for neglecting the context of pursuit, he himself completely neglects Kuhn's concepts of persuasion and translation, highly relevant for this issue.

it has shown a number of results, why can't they learn its initial results as well, and evaluate its initial pursuit worthiness? I'd like to suggest that the main reason why Kuhn rejects (or rather disregards) this possibility lies in his assumption that evaluating a new theory as worthy of pursuit amounts to actually pursuing it. We have already seen that his neglect of the communal pursuit worthiness leads him to this stance. And since, in his view, science can survive the change only by preserving diversity, that is, by allowing for different rivals to be simultaneously pursued, not all scientists should engage in a pursuit of a young theory. But once a theory shows further promise, scientists should gradually switch towards its acceptance. By reasoning in this way, Kuhn rejects the possibility of translation and persuasion in the early stages of theory pursuit in other to save the diversity of pursued theories.

However, once we have distinguished between individual and communal pursuit worthiness, it is easy to see that there is no reason why these processes would not be possible in the early stages of a pursuit.[10] This does not mean that every scientist is to learn any new theoretical candidate and evaluate whether it is initially worthy of pursuit. Rather, in order to be able to evaluate the initial pursuit worthiness of a new theory, a scientist must first learn about it. She must be able to step into the circle, and evaluate the merits and problems of the theory, even though she is currently working in a rivaling theory. If she does not learn about the new theory, then she is not competent in judging about its initial pursuit worthiness.

5.2.3 Conclusion

In view of the above quotes, we can conclude that while Kuhn did not explicitly consider the possibility of the communal notion of pursuit worthiness, he discussed certain aspects of the debates between scientists that can be helpful in this type of theory evaluation. He related these features only to the evaluation of successive pursuit worthiness, leaving the initial one to be assessed in terms of subjective, personal reasons. Nevertheless, we have argued that by neglecting to distinguish between individual and communal pursuit worthiness, Kuhn neglected the fact that the processes he suggests for the evaluation of successive pursuit worthiness are applicable to the evaluation of initial pursuit worthiness as well.

However, it is important to notice that even in the case of successive pursuit worthiness, an agreement regarding communal pursuit worthiness, that is, the question, which theories are worthy of pursuit for the given scientific community, is not guaranteed in view of the processes Kuhn suggests. Kuhn

[10] Hence, when Gerald Doppelt argues that in Kuhn's view "The arguments one paradigm presents to its rival may be powerful and sound; but they 'cannot be made logically or even probabilistically compelling for those who refuse to step into the circle'. (Kuhn 1996, p. 94)" (Doppelt 1978, p. 80), it is important to keep in mind that "refusing to step into the circle" does not amount to a cognitive or methodological necessity, but to refusing to take part in the process of translation. The fact that scientists indeed often avoid such a process speaks of their neglect of the requirements in the context of pursuit, rather than of a methodologically valid norm.

could have still argued that the differences among specific applications of the shared criteria may be such, that even the agreement about the communal pursuit worthiness may not always be possible.[11] Nevertheless, what is important here is that he suggests certain tools that can be helpful in reaching such an agreement. In addition to the process of translation, the idea of argumentative reasoning underlying the process of persuasion plays an essential role here. Moreover, it is this latter process that in Kuhn's view connects theory evaluation in the context of pursuit with the one in the context of acceptance, in terms of a gradual shift: "as argument piles on argument and as challenge after challenge is successfully met, only blind stubbornness can at the end account for continued resistance." (Kuhn 1996, p. 204). In the remainder of this section we will explore the significance of argumentative reasoning in theory change, pointed out by other philosophers who were inspired by Kuhnian insights.

5.3 Argumentative Approaches to Methodology – McMullin and Pera

5.3.1 McMullin on Meta-Theoretic Argumentation

While Kuhn finds argumentative reasoning to be essential for the debates regarding the pursuit worthiness or acceptance of scientific theories, he does not see it as a part of meta-theoretic changes, that is, changes in the epistemological and methodological standards upheld by scientists. With regard to the changes in the ways in which the set of shared epistemic standards are applied, or in the relative weights attached to them, Kuhn writes:

> ... many of these variations in value have been associated with particular changes in scientific theory. Though the experience of scientists provides no philosophical justification for the values they deploy (such justification would solve the problem of induction), those values are in part learned from that experience, and they evolve with it.(Kuhn 1977, p. 335)

In response to Kuhn, McMullin points out:

> This is to take the Hume-Popper challenge to induction far too seriously (unless, of course 'justification' were to be taken to mean definitive proof). The characteristic values guiding theory-choice are firmly rooted in the complex learning experience which is the history of science; this is their primary justification, and it is an adequate one. (McMullin 1982, p. 21)

[11] Therefore, incorporating Kuhn's stance on theory change into coherentist epistemology in order to preserve stronger notions of rationality may conflict with Kuhn's views – see the discussion in the Appendix of this thesis.

Hence, McMullin extends the use of argumentation also to the meta-theoretic discussions. According to him, the values involved in theory appraisal are instrumental, rather than ends in themselves: "They can be justified only by the extent to which they further the goals that science is taken to aim at." (McMullin 1984, p. 57). McMullin maintains that the rationality of science, the values used for the evaluation of theories can be philosophically justified:

> What philosophers of science have labored so long to show is that such values as fertility *are* an appropriate criterion of theory. Their arguments are in a broad sense logical or epistemological. ... What happens in philosophy of science reflects at the second level what happens in science itself. That is, it is empirically discovered in scientific practice that certain kinds of evaluative procedures or of epistemic demands (like the reproducibility of experimental results) are effective in bringing about the broadly-stated goals of science. (Ibid.)

The change of scientific goals is, therefore, based on reasons that reflect scientific practice. McMullin shows how the focus on prediction, which was characteristic for Babylonian astronomy, and on explanation, characteristic for Greek astronomy, conjoined into the complementary goals of the new science of the seventeenth century (ibid., p. 48). He argues that such changes occur on the basis of reasons internal to the scientific activity (p. 53). For instance, scientists may gradually realize that their ideal goals are in fact not achievable, and should be replaced with more realistic ones.[12]

If we perceive scientific changes in this way, argumentative reasoning becomes the cornerstone of scientific change and scientific methodology, both at the object level and meta level. In contrast to an algorithmic, rule-based approach to theory evaluation, which both Kuhn and McMullin reject (McMullin 1984, p. 56), the criteria underlying epistemic evaluation of theories are considered to function as values, presented in the form of arguments. One of the most detailed formulations of such an approach to theory evaluation has been given by Marcello Pera.

5.3.2 Pera's Dialectical Model of Science

Pera starts off from the shift that occurred in philosophy of science: from the traditional image of science, according to which science is perceived as certain, infallible, universal, and objective knowledge, guided by method that guarantees its aims, to a historically minded approach, which rejects the possibility of such a universal and precise method (Pera 1994, p. 1-5). Pera remarks that a common point of both of these views is an assumption that science without a method is not a cognitive and rational endeavor – an assumption which he dubs the *Cartesian syndrome* (p. 4).

[12] A similar point can be found in (Laudan 1984), where Laudan gives a number of reasons in view of which epistemic goals of science can be altered.

In contrast to many philosophers who belong to either the "methodological" or the "counter-methodological model" of science, Pera distinguishes Kuhn as a philosopher who managed to escape the Cartesian syndrome, and hence also the dilemma between methodological rationalism and irrationality: "he proposed a new way of understanding scientific rationality by replacing the old view based on method with a different one grounded in persuasive argumentation" (p. 10). Pera points out that this novelty introduced by Kuhn has been largely unrecognized, on the one hand, due to the fact that Kuhn had not developed a systematic and detailed account of his new approach, while on the other hand, the Cartesian syndrome prevented others from using the right categories needed to understand Kuhn's views on science.

In view of these insights, Pera takes a position "suggested but not fully developed by Kuhn" (p. 10). Aiming at finding a way out of the Cartesian dilemma, he proposes the *dialectical model of science*. In contrast to the two older models which view science as a game between two players – a researcher and nature, the dialectical model requires three players:

> a proposer who asks questions, nature that answers, and a community of competent interlocutors which, after a debate hinging on various factors, comes to an agreement upon what is to be taken as nature's official voice. In this model nature does not speak out alone. It only speaks *within* the debate and *through* the debate. (p. 11)

Pera defines his *dialectics* as the logic of convincing (or rhetorical) arguments, or as the logic of the rhetorical use of formal logics, or as a logic of debate (Pera 1994, p. 107-108), (Pera 2000, p. 60-61). In contrast to rhetoric which he defines as an act or practice of using convincing arguments, the task of dialectics is to *evaluate* such a practice. Pera suggests that the winner in a scientific debate is the one who refutes the arguments of his or her opponent, that is, who produces a convincing argument (Pera 2000, p. 59). A scientific argument (in a certain field of inquiry and for a certain purpose) is *convincing* if *i)* its premises are accepted and the factors it relies on are admitted by the community, in that field and for that function; *ii)* the rules of debate (the inference rules of formal logics and the material constraints ruling scientific debates) are considered pertinent by the community; and *iii)* it is valid according to such rules (p. 61). Hence, the opponents in scientific debates need to find certain areas if agreements which can serve as the basis of their argumentation:

> ... the contenders must first find, among the substantive factors of scientific dialectics, a minimal area of agreement (the *éndoxa*); on the basis of this area and their mutual confessions, they must then produce arguments with the aim of confuting each other. The debate is rationally over when the arguments of one party are stronger than those of the rival party. ... Those arguments that confute a rival on the basis of, or starting of with, a minimal area of agreement are stronger. Locating that area depends on the interlocutors'

> rhetorical abilities. Though it may not be easy if their opinions are far apart, it is not impossible, for even if a new opinion is radically innovative, there is no way it can alter or reject at the same time all the factors of scientific dialectics without being cast outside scientific tradition itself. (Pera 1994, p. 186-187)

Thus, even in the case of theories with radically different conceptual systems such a debate is possible:

> The obstacle we face in this case, that is, the problem of different references of the descriptive terms of the theories due to the different conceptual schemes they depend on, does not prevent a discussion between rival parties from taking place. As communication is possible even across such schemes, interlocutors may try to convince each other. It is enough for the parties to find a few shared premises to start with and then, by a sort of "rhetorical ascent", to go on step by step until one of them slowly yields while the other gets stronger and finally wins victory. (Pera 2000, p. 62-63)

Even though Pera does not make a distinction between theory evaluation in the context of pursuit and the one in the context of acceptance, it is easy to notice that his dialectics applies to both, the process of persuasion in the early stages of the development of a theory, as well as to the later stages, when one party wins over another. By focusing on scientific debates as "facilitators" of theory change, Pera shows in which ways Kuhnian translation and persuasion can actually be performed.

The aim of this brief overview of Pera's points is to show the significance he places on argumentative reasoning in theory change and theory evaluation. Moreover, instead of methodology based on strict rules of acceptance, rejection and preference, methodology is in Pera's model "absorbed by values" (Pera 1994, p. 117-118). In this respect, Pera takes a stance similar to Kuhn and McMullin, according to which, theory evaluation proceeds on the basis of values that are formulated in terms of arguments.

5.4 Conclusion: Modeling of Theory Evaluation in View of an Argumentative Approach to Methodology

Which conclusions can we make for the evaluation of pursuit worthiness in view of the argumentative shift in methodology?

In the case of informal modeling, such as our account of potential coherence, there are few important points to be made (some of which have already been mentioned in the previous two chapters):

1. the criteria of evaluation need to be understood as values, whose nature and weighting is to be interpreted in view of the specific historical context;

2. the historical context is to be presented in terms of scientific debates regarding the given theory;

3. both the arguments regarding the theory (object-level arguments), as well as the arguments regarding specific criteria used for theory evaluation (meta-level arguments) are to be taken into consideration.

This somewhat relativizes our account of pursuit worthiness in the sense that we have already mentioned at the end of Chapter 3. We have pointed out there that the main aim of our application of Bonjour's concept of coherence to the context of pursuit was to demonstrate that such an adaptation of an account of epistemic justification is possible. In case a specific historical context sets different criteria of theory evaluation, our framework might have to be adapted. We have presented this type of meta-justification of our framework in case of the continental drift debate in Chapter 4, where we have argued that our criteria fit the methodological standards of the geological community at the time. More precisely, we have shown that Drift was worthy of pursuit even in view of the standards upheld by its rivals.

However, in cases of more radical differences between rivaling theories, in which, for example, the explanatory power of a new theory is not even acknowledged by the standards of the rival, the evaluation is more complex. We have mentioned in Chapter 3 the case of Galileo's evidence obtained through the telescope, which is an example of such a problem. We have also mentioned in Section 5.3.1 the example of an even more radical difference in the evaluative standards of theory evaluation: the conflict between Galileian mathematical approach to astronomy and natural philosophy, with Aristotelian scientific practice for which mathematical demonstrations held no credibility with regard to material entities (see also (Biagioli 1993, p. 220)).[13] In such cases, meta-theoretic arguments, which, as we have seen, McMullin pointed out, need to be taken into account as well. The pursuit worthiness is then evaluated not only as the question: "Is this theory worthy of pursuit?", but also as the question: "Are there good epistemic reasons to allow for a different set of epistemic standards for theory evaluation?". Taking into account the epistemic aim of robustness of scientific knowledge (see Chapter 1) may be helpful in resolving this type of disputes, but what precisely the nature of such a meta-evaluation is, which criteria it consists of, how object-level arguments interplay with meta-level arguments, etc. requires a discussion of its own that goes beyond the scope of this thesis.[14] It is important to notice though that allowing for the changes in the criteria of theory evaluation does not mean that "anything goes". Rather, just like in the case of the evaluation of the pursuit worthiness of theories,

[13] I am indebted to Maarten Van Dyck for bringing Biagioli's work to my attention.

[14] (McMullin 1984) as well as (Laudan 1984, Chapter 3) provide interesting ideas of how this problem should be tackled from an epistemic point of view. In contrast, (Biagioli 1993) argues that both epistemic and non-epistemic factors were crucial for the seventeenth century mathematical shift in natural philosophy. Such an approach is challenged by Pera's (1994) discussion of Galileo's rhetoric, and the epistemic merit of his arguments.

what needs to be evaluated at the meta-level is the epistemic benefit of new epistemic standards.

In the case of formal modeling, the main challenge that the argumentative shift posed to formal approaches to theory evaluation is the question: how is formal modeling of theory choice at all possible if it should not be presented in terms of a set of pre-given methodological rules, functioning as an algorithm of theory choice? In the following chapter, we will propose a modeling that is based on arguments constituting scientific debates. The aim of this modeling will not be to offer a direct representation of our account of potential coherence. As has already been mentioned, our account could be modified to capture different accounts of epistemic justification. Hence, our primary concern will be the question, how to build a formal framework that can, in principal, model epistemic justification in both the context of pursuit and the context of acceptance. We will offer the basis of such a framework, which can be enhanced in different ways. Different enhancements can then be used to represent different epistemic and methodological criteria.

Chapter

6 Abstract Argumentation and Explanation Applied to Scientific Debates

✎ *This chapter is based on a paper with the same title, published in Synthese (Šešelja and Straßer 2011). The paper is a joint work with Christian Straßer. We are indebted to Erik Weber and two anonymous reviewers for valuable comments on a former draft of this paper.*

Summary Abstract argumentation has been shown to be a powerful tool within many fields such as artificial intelligence, logic and legal reasoning. In this chapter we enhance Dung's well-known abstract argumentation framework with explanatory capabilities. We show that an explanatory argumentation framework (EAF) obtained in this way is a useful tool for the modeling of scientific debates. On the one hand, EAFs allow for the representation of explanatory and justificatory arguments constituting rivaling scientific views. On the other hand, different procedures for selecting arguments, corresponding to different methodological and epistemic requirements of theory evaluation, can be formulated in view of our framework.

6.1 Introduction

Formal theories of argumentation have been extensively researched within the fields of artificial intelligence, philosophy, logic and computer science. One of the most influential accounts of argumentation is Dung's abstract argumentation framework (see (Dung 1993), (Dung 1995)). The significance of Dung's approach derives from the fact that it abstracts away from the nature of arguments and argumentation rules, which allows the user to focus on the interplay of arguments rather than on their specific structure. More precisely, an argumentation framework (AF) consists of a set of arguments $\mathcal{A}$, which are taken to be abstract entities represented by alphabetical letters, and the binary (so-called *attack*) relation $\rightarrow$ defined over this set. AFs are used to select sets of

arguments from $\mathcal{A}$ that satisfy certain standards of acceptability. Selection criteria are defined in order to explicate these standards: for instance selected sets of arguments are supposed to be non-conflicting and to be able to defend themselves against all argumentative attacks.[1] An extensive research on abstract argumentation has shown that such systems are capable of formalizing various approaches to nonmonotonic reasoning in the fields of artificial intelligence, logic programming and human reasoning. The fruitfulness of Dung's framework stems not only from its abstract character, but also from the fact that it is easily enhanceable with additional properties and useful in different application contexts.[2]

In this chapter we will enhance AFs with explanatory features. The aim of this enhancement, which we will call an Explanatory Argumentation Framework (EAF), is, on the one hand, to equip AFs with tools that can model explanatory reasoning, and on the other hand, to demonstrate that abstract argumentation provides a useful formal framework for the modeling of scientific debates. The basic idea of our enhancement is to introduce to AFs a set of explananda and an explanatory relation. This will allow us to express certain notions, such as explanatory power and explanatory depth, in terms of our framework. Moreover, we will show that EAFs allow for a comparison of different sets of arguments in view of their explanatory virtues. Taking into account that scientific explanation is one of the key constituents of scientific reasoning, EAFs will turn out to be a handy modeling tool in fields dealing with the reconstruction and the modeling of scientific debates, such as the philosophy of science. To this end we will offer a set of criteria which are useful for the demarcation of rivaling scientific views in terms of arguments, as well as for an evaluation of such views in terms of their argumentative and explanatory properties. As a result we will be able to formulate new selection criteria, suitable for the modeling of argumentation and explanation in a scientific context. Finally, we will show that our approach may be embedded or linked to the argumentative shift in methodology that is associated with scholars such as Marcello Pera (see the discussion in the previous chapter) and Marcelo Dascal. Since a number of enhancements developed for AFs can also be applied to EAFs, we will suggest that in this way abstract argumentation can provide an even more refined and more realistic modeling of scientific debates.

The chapter is structured as follows. We begin in Section 6.2 by explicating the close relation between argumentation and explanation, on the basis of which we will motivate the significance as well as the structure of our framework. In Section 6.3 we introduce the basic notions of abstract argumentation. In Section 6.4 we present EAFs. In Section 6.5 we informally introduce criteria and selection procedures that allow for a more realistic representation of scientific reasoning than the standard selections offered within Dung's abstract

[1] We give a formal account of this and other standard selection criteria in Section 6.3.

[2] For the enhancements that have been developed for AFs see Section 6.7.3. As for the different application fields, for instance, AFs have been used for an improved account of default reasoning (Bondarenko et al. 1997), (Dung and Son 1996), as well as for multi-agent systems (Coste-Marquis et al. 2007), (Bench-Capon 2003).

argumentation framework. In Section 6.6 we formally explicate the explanatory properties that have been previously introduced. Section 6.7 offers a discussion on some additional questions concerning the virtues of our framework. We show here that EAFs reflect some of the key ideas underlying rhetorically minded approaches to scientific rationality, and we point out the novelties of our framework, as well as possible enhancements of it. Section 6.8 concludes the chapter.

6.2 Argumentation and Explanation

Explanation and argumentation have been studied in philosophy of science, epistemology and logic. While some authors have discussed the two in close relation, others have pointed out the need to distinguish them as two different processes of reasoning. In this section we will explicate the relation of argumentation and explanation in our framework and situate it within the broader context of the discussion on this matter.

6.2.1 The Goal-Directed Perspective

One way to look at the problem of distinguishing argumentation and explanation is to explicate what it is that explanations try to achieve. Hughes states that

> the purpose of an explanation is to show *why and how* some phenomenon occurred or some event happened; the purpose of an argument is to show *that* some view or statement is correct or true. Explanations are appropriate when the event in question is taken for granted, and we are seeking *to understand* why it occurred. Arguments are appropriate when we want to show that something is true, usually when there is some possibility of disagreement about its correctness. (Hughes 1992, p. 76, italics added)

Thus, the goal of an explanation is to reach an understanding of the *why* or *how* something occurred, depending on the type of explanation. The occurrence itself is thereby taken for granted.

The quotation above suggests even more, namely that explanations are distinguished from arguments due to the different types of goals that the respective notions achieve. In contrast to explanations, arguments are justificatory, they show *that* something is the case and not why or how. Thus, the quotation suggests a clear distinction between arguments and explanations. However, we will subscribe in the following to the view that justificatory arguments are a certain subclass of arguments, and that explanations (in a strict sense) should be conceived of as a certain type of arguments as well.

6.2.2 Explanations as Arguments

The view that explanations are arguments has a long history. According to Hempel's covering law model of explanation, which is considered to be one of the origins of the contemporary study of explanation,[3] an explanation is an argument in which a sentence describing a phenomenon to be explained is derived from the class of those sentences which are adduced to account for this phenomenon, and which contain at least one law of nature (Hempel 1965, p. 247). A similar view on explanations as arguments or argument patterns can be found in unificationist accounts of explanation (e.g. see (Kitcher 1981b), (Weber 1999)). Moreover, the view that some arguments have an explanatory function is not foreign to the literature on argumentation either (e.g. see (Pera 1994) p. 110, (Pera 2000) p. 57).

In order to see which type of arguments explanations are we should first of all analyze the notion of an argument a bit more. Mayes in his (Mayes 2000) distinguishes between two meanings of this term: a formal and an evidentiary one. In a broader, *formal sense*, an argument is a finite sequence of propositions (called premises) followed by a proposition (called conclusion), in which the premises are intended (or taken) to entail the conclusion (ibid., p. 363). In a narrower, *evidentiary sense*, we are speaking of a specific type of formal arguments, namely those in which premises provide a rational justification for believing the conclusion (ibid., p. 364). This is the sense in which Hughes uses this term and what we have called justificatory arguments. However, as Mayes points out, beside justificatory arguments there is another type of formal arguments: explanatory ones or simply, explanations. The basic difference between these two types of arguments, as we have already seen, is that while justificatory ones aim at justifying *that* something is the case, explanatory ones aim at answering the question *why* (or *how*) something is the case.[4]

In this sense, an explanation is a formal argument consisting of an explanans and an explanandum, where the former one offers the causes or the governing law of the latter one and thus provides a better understanding of it. That is, premises of an explanatory argument represent an explanans from which a conclusion, representing an explanandum, can be inferred on the basis of a certain inference relation (such as deduction, induction, etc.).

[3] The contemporary study of explanation is usually seen as originating in (Hempel and Oppenheim 1948), which was further developed in (Hempel 1965).

[4] It is important to notice that sometimes we can determine whether a given argument is justificatory or explanatory only by taking into account the given context, which reveals the intention of the speaker. For example, an elliptically expressed argument "Shops are closed today because it's a public holiday." – could in one context be an explanation given in reply to the question "Why are shops closed today?", where the fact that shops are closed is taken for granted for both participants involved in the conversation. In some other context though, the same argument could be expressed as a justification of the fact that shops are closed, where this fact is doubted by one of the participants.

6.2.3 The Processual Character of Explanations

Let us in the following put more emphasis on the notion of understanding. By offering an explanation to an explainee, the explainer tries to make the explainee understand why/how/etc. the explanandum occurred. However, nothing guarantees that after offering an explanatory argument, the explainee has actually reached the point of understanding. Often an explanatory argument needs to be complemented by a dialogical process that clarifies certain open questions or doubts on part of the explainee. Thus, we can perceive explanations in a broader sense to be an argumentative process aiming at the explainee's understanding of the given phenomenon. Such a processual character of explanations has been emphasized, for example, by Schurz who speaks of 'explanatory episodes', characterized as relations between two cognitive systems communicating with each other in order to achieve a better understanding of the phenomena in question (Schurz 1991).

An explanatory episode is considered to be a process which includes not only explanatory arguments but may also include justificatory arguments, where the task of the latter ones is to further substantiate the former. Upon hearing an explanation, the explainee may request further clarification and may express his doubt for some of the arguments by either challenging (some of) them with counter-arguments or by requesting further clarification. Consequently, the explainer may have to justify claims constituting her explanation. Thus, arguing is often a constitutive part of an explanatory process not only because the explainer may wish to explicate and strengthen her claims, but also because the validity of some of them may be brought into question in case the explainee does not find them sufficiently accurate, clarified, understandable, etc. Consequently, explanatory reasoning does not have to result only in knowledge accumulation, but may sometimes also include a revision and thus contraction of the knowledge base of the explainee (see (Schurz 1991)).

Argumentation is thus a constitutive feature of explanatory reasoning. Together with Mayes we can say that, "until an explanatory hypothesis has been independently established through argument, it lacks the power to support anything at all", and the other way around, "until a justified belief has been adequately explained it lacks the power to support anything at all" (Mayes 2000, p. 375). Let us take a closer look at Mayes' description of such an interactive relation:

> Explanation is a process that is triggered by a certain kind of input, viz., a surprising fact, a salient feature of our environment that we have somehow failed to predict. (E.g., the car wont start). ... Explaining a fact involves the formation of a causal hypothesis (The battery is dead.). This possible cause is the output of the explanatory process. But for any given fact there will always be a number of possible causes. Hence, the process of explanation will be useful as a way of gaining predictive control over our environment only if it is supported by another process whose function is to

> determine which, if any, of the possible causes should be accepted. (ibid., p. 378).

6.2.4 Explanation and Argumentation in the Context of Scientific Reasoning

In this chapter we will primarily focus on the modeling of scientific explanations, or more precisely, scientific explanatory reasoning. In addition to the dynamics of explanation and argumentation which has to be taken into account in such a modeling, it is important to notice that a bilateral relation, involving one explainer and one explainee, is not the only possible situation in an explanatory process. This is, for instance, the case in scientific contexts where a number of scientists can participate in a discussion on a certain explanatory issue. In such situations, the explanation proposed by one scientist (or a group of scientists) undergoes a critical assessment by the other members of the given scientific community. Moreover, different scientists may offer different, mutually rivaling explanations. As a result, arguments used in explanatory reasoning will be open for criticism in terms of counterarguments, while explanations will be open for a comparison with other alternative explanations.

Thus, on the basis of the points presented in this section we can conclude that an appropriate modeling of scientific explanatory reasoning should allow for the following three properties:[5]

1. a dynamic view on explanatory reasoning, involving both justificatory and explanatory arguments;
2. the possibility of expressing criticism in terms of counterarguments and alternative explanations;
3. the possibility of multiple participants in an explanatory process.

In this chapter we will offer a framework that can satisfy all three of these requirements. First of all, rooting our framework in Dung's account of abstract argumentation allows for an abstract notion of an argument, which can be seen as corresponding to an argument in a formal sense. Consequently, both justificatory and explanatory arguments can be represented as argumentative letters in general. Second, the dynamics of abstract argumentation, based on the attack relation between arguments, allows for a modeling of counterarguments and alternative explanations. Finally, as we will demonstrate in our examples, an abstract argumentation system allows for the input from multiple parties to be represented in an explanatory process, which further contributes to its fitness for the modeling of scientific explanatory reasoning and scientific debates.

Before we introduce our framework, let us give a summary of the main concepts of Dung's abstract argumentation.

[5] Even though these properties are important for the modeling of scientific explanatory reasoning, they are not restricted to it. Similar kind of requirements may be posed on the modeling of other explanatory contexts such as e.g. expert systems.

6.3 Abstract Argumentation

Let us first have a look at the classical definition of argument systems introduced by Dung in (Dung 1995). We have a set of arguments and an attack relation between them. The abstractness of the framework concerns both elements. On the one hand, we do not reveal the concrete structure of the given arguments, but represent them by abstract letters. On the other hand, we do not reveal the concrete nature of the attack relation.

Definition 6.3.1. An *argumentation system (AF)* is a pair $(\mathcal{A}, \rightarrow)$ where $\mathcal{A}$ is a set of arguments, and $\rightarrow \subseteq \mathcal{A} \times \mathcal{A}$ is a relation between arguments. The expression $a \rightarrow b$ is pronounced as "a attacks b" and $\rightarrow$ is called the *attack relation*.

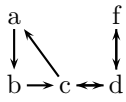
The central notion of AFs is acceptability. We are interested in selecting sets of arguments, let us call them *A-sets*, which satisfy criteria of acceptability.[6] For example, the selected arguments should be at least conflict-free or should be able to defend themselves from all the attacks by other arguments. Applied to scientific discourse, an A-set represents a collection of arguments that satisfies a certain, for instance, methodological virtue. The following definitions introduce the standard selection criteria for A-sets.[7]

Definition 6.3.2. Given an argumentation framework (AF) $(\mathcal{A}, \rightarrow)$ and an A-set $A \subseteq \mathcal{A}$ we define:

(i) A defends the argument a iff every attacker of a is attacked by a member of A.
(ii) A is *conflict-free* iff no argument in A attacks an argument in A.
(iii) A is said to be *defended* if it is conflict-free and every argument in A is defended by A.[8]
(iv) We call maximal (w.r.t. $\subseteq$) defended A-sets *preferred A-sets*.

Example 6.3.1.

We will demonstrate the concepts just introduced with the attack-diagram to the right. The table lists the A-sets belonging to selections based on the different criteria:



conflict-free	defended	preferred
$\varnothing, \{a\}, \{b\}, \{c\}, \{d\}, \{f\}, \{a,d\}, \{a,f\}, \{b,d\}, \{b,f\}, \{c,f\}$	$\varnothing, \{d\}, \{f\}, \{a,d\}$	$\{a,d\}, \{f\}$

[6] These were introduced by Dung as so-called "extensions".

[7] Many other selection criteria for A-sets have been proposed in the literature (such as being "stable" and "complete" in (Bondarenko et al. 1997), being "semi-stable" in (Caminada 2006), being "ideal" in (Dung et al. 2007), etc.). In order to make the technical level of the chapter not too involving we stick to the selection criteria introduced in Definition 6.3.2. Generalizations of our framework for other selection criteria are straight-forward.

[8] Defended A-sets are also often labeled "admissible".

6.4 Enriching Abstract Argumentation with Explanations

In this section we will define explanatory argumentation frameworks (EAFs) and some basic notions that can be expressed by them.

6.4.1 Explanatory Argumentation Frameworks (EAFs)

In order to equip argumentation frameworks with explanatory capabilities we extend them with the following three elements:

1. A *set of explananda* $\mathcal{X}$: we interpret elements of the set $\mathcal{X}$ as statements describing a state of affairs which is considered to be requiring an explanation by all the parties involved in the given dispute or which is within the explanatory scope of a given discipline. This could be a certain natural or social phenomenon, an experimental result, etc. In accordance with the standard view on explanations which take the explanandum as indisputable in character (in contrast to the conclusions of evidentiary arguments), we assume that the set of explananda consists of facts which are considered to be indisputable in the given field. For example, an explanandum can be a description of, or a reference to a certain observation or an experimental result.[9]

2. The second element we need to introduce is an *explanatory relation* $\dashrightarrow$ which holds between:

(a) an argument and an explanandum, i.e., $\dashrightarrow \subseteq \mathcal{A} \times \mathcal{X}$ where $\mathcal{A}$ is the set of arguments of a given AF and $\mathcal{X}$ is the set of explananda;
(b) between two arguments, i.e., $\dashrightarrow \subseteq \mathcal{A} \times \mathcal{A}$.

Where $a \in \mathcal{A}$ and $x \in \mathcal{A} \cup \mathcal{X}$ we designate "$a \dashrightarrow x$" as "a explains x". While the explanatory relation between an argument and an explanandum links phenomena requiring explanation with the reasons which should allow for their better understanding, the explanatory relation between arguments themselves allows for explanations to be deepened. In other words, argument b can be used to explain one of the premises of argument a (which may itself be used to explain explanandum e) or the link between the premises and the conclusion. The former case corresponds to Thagard's idea of a deepening of scientific explanations or Bermúdez's notion of a vertical explanation.[10] The latter case can occur in explanatory situations typical for everyday language, didactic situations or oral disputes, in which arguments are usually expressed in an elliptic manner so that the link between premises and the conclusion might not be sufficiently clear. In accordance with the abstract character of abstract argumentation frameworks, we treat the explanatory relation in an abstract manner as well.

3. We introduce the third element that simplifies the modeling of scientific debates while making it at the same time more accurate. Sometimes two

[9] Nevertheless, sometimes there are disputes on what is to count as a *valid* or *important* explanandum in a given scientific field. For the possibility of enhancing our framework so that it can allow for such disputes, see Section 6.7.3.

[10] We will present both notions in Section 6.6.

arguments a and b are based on incompatible presuppositions or premises. It is important to notice that this does not necessarily indicate that a attacks b or vice versa. This is often the case with alternative explanations of a certain phenomenon. For instance, some geologists in the first half of the twentieth century explained the origin of mountains by the idea of continental drift (g), while some other geologists explained it by the thesis of the earth's contraction (c). Although g and c were clearly incompatible (see Example 6.4.1, Section 6.4.2), g in itself was not sufficient to attack c (and vice versa): just naming an alternative explanation is not considered as a counter-argument in a scientific debate. And indeed, counter-arguments against both ideas were established on independent grounds. For instance, contractionists attacked the theory of drifters by pointing out that it cannot account for a mechanism that would enable continents to plough through the dense seafloor. In order to model such incompatibilities between arguments we introduce another relation, the *incompatibility relation* $\sim$.

In conclusion, we define:

Definition 6.4.1. An Explanatory Argumentation Framework (EAF) is a tuple $\langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow, \sim \rangle$, where $\langle \mathcal{A}, \rightarrow \rangle$ is an AF, $\mathcal{X}$ is a set of *explananda*, $\dashrightarrow \subseteq (\mathcal{A} \times \mathcal{X}) \cup (\mathcal{A} \times \mathcal{A})$, and $\sim \subseteq \mathcal{A} \times \mathcal{A}$ is a symmetric relation.

We call $\dashrightarrow$ the *explanatory relation* and the elements of $\dashrightarrow$ *atomic explanations*. The elements of $\mathcal{X}$ are denoted by $e, e_1, e_2, ...$, and the elements of $\mathcal{A}$ by $a, b, c, d, f, g, ...$ Moreover, $\sim$ is the *incompatibility relation* and in case $a \sim b$, a and b are said to be incompatible.

Concerning the selection criteria introduced in Definition 6.3.2 little adjustment is needed. The only change concerns the notion of conflict-freeness: In the remainder of the chapter we call an A-set A *conflict-free* iff no argument in A attacks *or is incompatible* with an argument in A. As before, A is said to be *defended* if it is conflict-free and every argument in A is defended by A.

6.4.2 Basic Definitions

In order to introduce some basic notions it is useful to first define some basic graph-theoretic concepts.

Definition 6.4.2. A *directed graph (digraph)* is an ordered pair $G = \langle V, \rightsquigarrow \rangle$ where V is a set and $\rightsquigarrow$ is a binary relation on V, $\rightsquigarrow \subseteq V \times V$. The elements of V are called vertices and the elements of $\rightsquigarrow$ are called arrows. $G' = \langle V', \rightsquigarrow' \rangle$ is a *sub-graph* of G iff $V' \subseteq V$, $\rightsquigarrow' \subseteq V' \times V'$ and $\rightsquigarrow' \subseteq \rightsquigarrow$. G' is a *proper* sub-graph of G iff it is a sub-graph of G and $V' \subset V$ or $\rightsquigarrow' \subset \rightsquigarrow$. We say that there is a *path from x_1 to x_n in G* iff there are $x_1, \ldots, x_n \in V$ for which $(x_i, x_{i+1}) \in \rightsquigarrow$ for all $1 \leq i < n$. G is *circular* iff there is an $x \in V$ for which there is a path from x to x. Where $V' \subseteq V$, we define $\rightsquigarrow_{V'} =_{\mathrm{df}} \{(x, y) \in \rightsquigarrow \mid x, y \in V'\}$.

We now introduce definitions that characterize explanations in EAFs.

Definition 6.4.3. Let $\mathsf{A} = \langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow, \sim \rangle$ be an EAF, $e \in \mathcal{X}$, and $a \in \mathcal{A}$.

(i) We call a sub-graph $X = \langle A, \dashrightarrow_A \rangle$ of $\langle \mathcal{A}, \dashrightarrow_{\mathcal{A}} \rangle$ an *explanation of* e iff there is a unique argument $a \in A$ such that (i) $a \dashrightarrow e$ and (ii) there is a path in X from every $a' \in A \setminus \{a\}$ to a.
We say that the explanation X is circular if X is a circular graph.
We use the following writing conventions: In the case that an explanation X only consists of a path $\langle P, \dashrightarrow_P \rangle$, where $P = \{a_1, \ldots, a_n\}$ and $\dashrightarrow_P = \{(a_{i+1}, a_i) \mid 1 \leq i < n\}$, we abbreviate X by $\langle a_1, \ldots, a_n \rangle$. We sometimes write $X[e]$ for X in order to indicate that X explains e.

(ii) An explanation $\langle A, \dashrightarrow_A \rangle$ is *conflict-free* iff A is conflict-free.

(iii) An explanation $X[e]$ is *deeper* than an explanation $X'[e]$ iff X' is a proper sub-graph of X. We write $X' \prec X$. We say that X' is a *sub-explanation* of X.

(iv) $X[e]$ and $X'[e]$ are *alternative explanations* of e iff neither $X \prec X'$ nor $X' \prec X$.

(v) An explanation $\langle B, \dashrightarrow_B \rangle$ is *offered by an A-set A* iff $B \subseteq A$. We define the set of all explananda for which A offers an explanation by

$$\epsilon(A) =_{\text{df}} \{e \in \mathcal{X} \mid \text{there is an explanation } X[e] \text{ offered by } A\}.$$

Example 6.4.1. We give an example of a scientific debate, on the basis of which we can clarify the notions that have so far been introduced, and which will serve to show why and how our framework can be useful in an evaluation of scientific theories. The following arguments, which correspond to the EAF given in Figure 6.1, are central to (though they do not exhaust) the discussion in the geological sciences around the 1920's, which we have seen in Chapter 4.

Scientific debates are often very technical and complex. Hence, in order to follow them a scholar has to be sufficiently familiar with the involved topics. We chose this example since the technical complexity of the given arguments allows for a representation that is understandable and transparent also for scholars that are not already familiar with the subtleties of the research in geology at that time and it is thus ideal as a running example for demonstrating our framework. We begin with an excerpt of the arguments given by drifters and contractionists, which will be further extended at a later point in this chapter (see Example 6.5.1). The explananda are as follows:[11]

$\mathbf{e_1}$ (*fossils*) Similar kinds of fossils were found on different continents.[12]

$\mathbf{e_2}$ (*orogeny*) There are mountains and mountain chains on continents.

$\mathbf{e_3}$ (*glaciation*) There is an evidence of glaciation which took place in the late Paleozoic in the southern continents (the so-called Southern Glaciation

[11] Even though we will, for the sake of simplicity, focus on some arguments exchanged between the drifters and the contractionists, it is important to notice that the permanentist side could easily be included in our example, and that EAFs are suitable for the modeling of any number of parties involved in an explanatory process.

[12] This is a simplified version of the actual explanandum, which states a peculiar distribution of Cambrian trilobites – fossil arthropods that lived 500 to 600 million years ago (see (Gould 1977)); we will make similar simplifications of other explananda and arguments constituting this example in order to avoid burdening the reader with too many technical details.

or the late Paleozoic glaciation).

The following arguments were offered:

a (*land bridges*) In the past, the continents were apart like nowadays, but connected by land bridges. This is how different species of flora and fauna were distributed to different parts of the world.

b (*no land bridges nowadays*) The hypothesis of the land bridges is not plausible since it is not clear how such land bridges would have disappeared throughout the history.

c (*contraction*) Vertical displacements of the otherwise unmovable earth's crust result from the contraction of the earth, which causes shrinking and lateral compression in the crust. That is why some rocks (such as mountains) became elevated while some others (such as the land bridges) subsided into the ocean.

d (*cooling*) The earth is contracting due to its cooling.

f (*drift-paleontology*) Continents were once connected into a super-continent, before they drifted away from each other. Different species of flora and fauna were distributed over different continents in this way.

g (*drift-orogeny*) Drifting of continents results in the leading edge of the continent being compressed and folded upwards due to the resistance of the seafloor. Consequently, mountains are being formed along the leading coastlines of a drifting continent, or result from two continents colliding against each other.

h (*drift-glaciation*) The nowadays southern continents were once a part of a super-continent, and positioned more in the north. That is why glaciation could occur on them in Paleozoic, before they drifted to the south.

i (*drift*) The earth consists of concentric shells, the density of which increases from the crust to the core, so that the continents float on and extend into the ocean floors. This is why the continents, pulled by a particular (currently unknown) force, could drift away from their original locations where they once formed a super-continent.

Let us focus on maximal conflict-free sets of arguments that are able to defend themselves in order to gain a first approximative representation of the rivaling scientific views.[13] Hence, we are interested in *preferred* A-sets. In this example we have two such A-sets: $A_1 = \{f, h, g, i\}$ and $A_2 = \{a, c, d\}$, corresponding to the two represented rivaling views in geology: Drift and contractionism. Next, we have two atomic explanations of e_1: a and f, two atomic explanations of e_2: c and g, and one atomic explanation of e_3: h. Each of them is a sub-explanation of the following explanations, resp.: $X_1[e_1] = \langle a \rangle$, $X_2[e_1] = \langle f, i \rangle$, $X_3[e_2] = \langle c, d \rangle$, $X_4[e_2] = \langle g, i \rangle$, and $X_5[e_3] = \langle h, i \rangle$ all of which are conflict-free and non-circular. By Definition 6.4.3iv, $\langle g, i \rangle$ is deeper than $\langle g \rangle$ alone. Explanations X_2, X_4 and X_5 are offered by the drifters, i.e. A_1, and explanations X_1 and X_3 are offered by the contractionists, i.e. A_2. Notice that

[13] We will offer more realistic selection procedures for the representation of scientific views in Section 6.5.

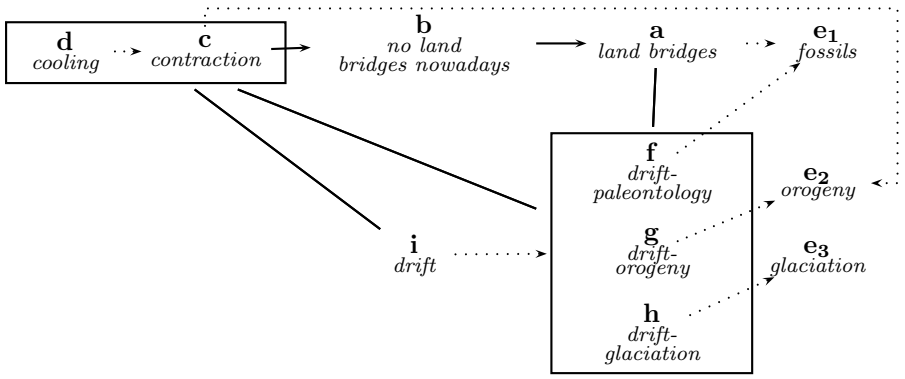


Figure 6.1: The EAF of Example 6.4.1. Solid arrows represent the attack relation, dotted arrows represent the explanatory relation, and solid lines represent the incompatibility relation. Solid lines from the box around arguments f, g and h to argument a and the box around d and c indicate that all three arguments are incompatible with a, d and c. The explanatory arrow from i to the box around f, g and h indicates that each of the three arguments is explained by i.

the two preferred A-sets, A_1 and A_2, offer explanations for different sets of explananda: while A_2 offers explanations for e_1 and e_2, A_1 offers explanations for e_1, e_2 and e_3. Hence, while the two A-sets are in view of their argumentative properties equivalent (i.e. they are both maximally defended and conflict-free), their explanatory power is different. Since a difference in the explanatory power of A-sets can play an important role in the evaluation of scientific theories, we will introduce criteria for selecting A-sets in view of their explanatory features in Sections 6.5 and 6.6.

This example also demonstrates the usefulness of our incompatibility relation. We take the contractionists' arguments a, c and d to be incompatible with all of the explanations given in view of the Drift (f,g,h and i) since they assume mutually incompatible explanatory mechanisms.[14] Obviously, argument d refers to the level of cooling that can account for the level of contraction needed to explain the formation of mountains (e_2), and not to a more moderate version of cooling and contracting, which would not be able to explain such a phenomenon and which would be compatible with the Drift. Notice that without our incompatibility relation, the arguments of the two sides would have to be modeled either as formally unrelated in terms of EAFs or as related in terms of bidirectional attacks (in place of the incompatibility relations). However, in the first case, it would be impossible to distinguish between the two

[14]For example, one of the reasons for this incompatibility (which for the sake of simplicity we have kept out of our examples) lies in the fact that Drift relied on the principle of isostasy, which implied that continents and ocean floors had to be different either in structure or in composition, and which conflicted with the contractionists' idea of interchangeability of oceans and continents (Oreskes 1999, p. 21-55).

rivaling scientific views, since for instance $\{a, c, d, f, g, h, i\}$ would be a conflict-free A-set. The second option would allow for the distinction between the rivaling views, but it would have some other implausible results. For example, argument b would be taken as defended from c by any of the Drift arguments f, g, h, i merely due to the fact that an alternative explanation has been proposed, which would be counterintuitive.

6.5 Towards a More Realistic Modeling of Scientific Debates

6.5.1 Criteria for the Modeling of Scientific Debates

As we have seen in Section 6.4, certain criteria for A-sets (such as being conflict-free or defended) are useful for the representation of opposing views in scientific debates. However, for a more realistic modeling of scientific views and their evaluation, we need to add few more criteria and modify some of those that have already been introduced. We will show that some of the key epistemic values relevant in the evaluation of scientific theories can be expressed in terms of our framework. On the basis of them, we will be able to formulate selection types for A-sets that reflect certain methodological and epistemic preferences scientists or philosophers may have when evaluating theories in view of the available arguments. We will propose two procedures for such selections, which are more apt for this purpose than the standard criteria introduced in Section 6.3.

It is important to mention though that it is beyond the scope of this thesis to finally settle the question, which criteria (and combinations thereof) most adequately capture the methodological and epistemic standards used in theory evaluation, either descriptively or normatively. Many new criteria have been studied since Dung developed his AFs, which refine and optimize the first generation of selection criteria in many ways.[15] It is a task left for future research to clarify which (combinations of) criteria are the most suitable for the modeling of the notions of acceptability underlying theory choice. What we want to present here is rather a general directive for how this research may proceed and in which way notions developed in terms of EAFs can be useful for this task.

The criterion of conflict-freeness, introduced in the previous section, is a minimal requirement that should be satisfied by A-sets representing a given scientific view. In view of this criterion we can then distinguish between mutually rivaling scientific views. Another epistemic standard significant in the evaluation of scientific theories is their explanatory power.

Explanatory power. The explanatory power of an A-set – usually also referred to as explanatory *scope* or explanatory *breadth* – is given by the set of explananda which are explained by its constituting arguments. We are interested in sufficiently explanatory powerful (conflict-free) A-sets. There are two

[15] See Footnote 7.

ways of comparing the explanatory power that may be relevant in the assessment of scientific theories. On the one hand, if an A-set has a (clearly) smaller explanatory scope compared to another A-set, then it is usually considered to be a suboptimal candidate in the context of theory acceptance (other things being equal). On the other hand, when we are evaluating whether a new scientific theory is worthy of pursuit, it may be enough for an A-set to have some novel explanations, i.e. to explain certain explananda that are not explained by any alternative A-set. Therefore it will be useful to introduce two ways of comparing the explanatory power of A-sets: on the one hand in a *quantitative* sense and on the other hand in a *qualitative* sense. We will discuss them in more detail in Section 6.6.1. Since the aim of this section is to present the main idea underlying our new criteria, we will use only a simplified version of the latter comparison type which we informally define as follows:

We say that an A-set A_1 is explanatory more powerful than an A-set A_2 iff the set of explananda for which the arguments in A_1 offer an explanation is a proper super-set of the set of explananda for which A_2 offers an explanation (i.e., $\epsilon(A_2) \subset \epsilon(A_1)$).

Example 6.5.1. In order to get a more accurate picture of the discussion in geological sciences presented in Example 6.4.1, we extend it with some additional arguments. The EAF corresponding to the example is given in Figure 6.2.

j (*mechanism-problem*) It is not at all clear how the continental drift can occur, since continents cannot simply plough through the dense seafloor.

k (*radioactivity*) Due to the discovery of radioactive material in the earth's crust, which produces heat when decaying, we can claim that the earth cannot be cooling, at least not to such an extent that would account for the origin of higher mountain chains.

l (*why contracting?*) It is not plausible to assume that the earth is contracting unless we know the causes of such a process, and no such cause seems to exist.

The most explanatory conflict-free A-sets are various super-sets of $A_1 = \{f, g, h\}$, e.g. $A_2 = \{f, g, h, i, l\}, A_3 = \{b, f, g, h, k, l\}, A_4 = \{b, f, g, h\}$, etc. Note that any conflict-free super-set of $\{a, c\}$ (i.e. A-sets representing contractionism) only explains $\{e_1, e_2\}$ and is hence explanatory weaker than $A_1, \ldots, A_4$. Indeed, given the arguments introduced so far, the Drift offers a broader explanatory scope than Contractionism.

However, it is important to notice that in this example none of the most explanatory A-sets, such as $A_1, \ldots, A_4$, are defended in a strict sense: after all neither of these sets is able to defend itself from the attack by j. The only preferred A-set is $A_1' = \{b, j, k, l\}$. Nevertheless, this set has no explanatory power with respect to the given explananda. Note that the two preferred A-sets from Example 6.4.1 – $\{f, g, h, i\}$ and $\{a, c, d\}$ – both offer a greater explanatory power than A_1'. However, they are not anymore selected, since they are not able to defend themselves from all the attacks. This situation is not atypical in science since many theories that are accepted or pursued are confronted

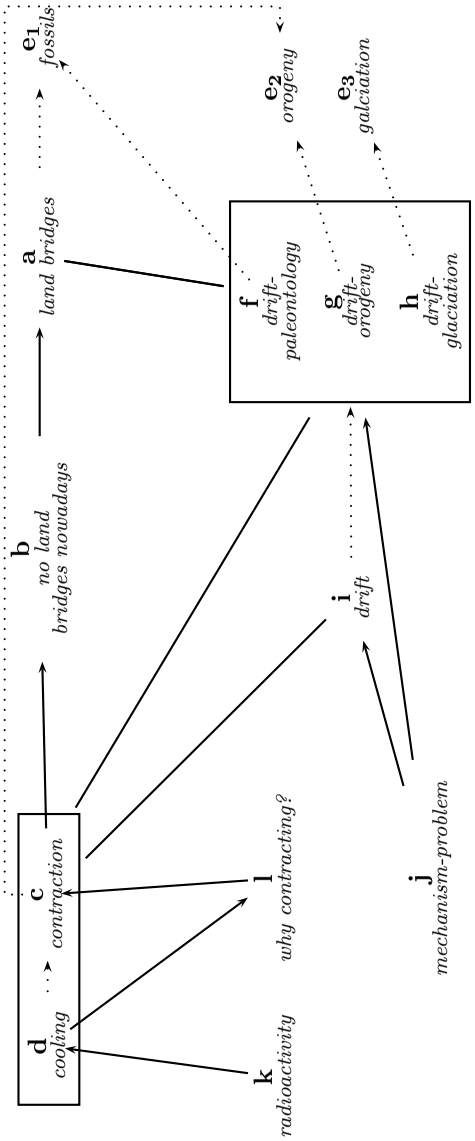


Figure 6.2: The EAF from Example 6.5.1.

with objections and criticisms of various kinds, from which they cannot always immediately be defended. This is especially so in the context of pursuit where we are primarily interested in a theory that can offer explanations for certain phenomena in spite of having some open problems. For example, during the confrontation of the above mentioned geological theories, the fact that none of them was resistant against criticism was not a sufficient reason for abandoning their further pursuit. Hence, in certain epistemic contexts we might wish to lower our standard of defense, that requires from an A-set to defend itself from all attacks.

Weakening the Standard of Defense. The idea of this weakening is to say that an A-set A_1 is more defended than another A-set A_2 iff A_2 is attacked by more arguments against which it cannot defend itself than A_1.[16]

Let us take a look at the most explanatory A-sets from Example 6.5.1. A_2 is attacked by both j and d. Similarly, A_4 is attacked by j and c. The A-sets A_3, $A_5 = \{f, g, h, i\}$, and $A_6 = \{f, g, h, i, k\}$ fair better: they are only attacked by j. The latter sets belong to the most defended of the most explanatory A-sets.

Explanatory Depth. In addition to comparing sets of arguments in view of their explanatory power, we can also compare them in view of their *explanatory depth*. For instance, both $\{f, g, h\}$ and $\{f, g, h, i\}$ have the same explanatory breadth since both offer explanations for all the shared explananda e_1, e_2 and e_3. However, $\{f, g, h, i\}$ is explanatory deeper than $\{f, g, h\}$ since i explanatory deepens arguments f, g and h. Note that the latter are not shared explananda but theory-internal parts of the Drift. Since we are interested in representing a scientific view as consisting not only of the arguments that directly explain the shared explananda (e.g. $\{f, g, h\}$), but also the arguments that explanatory deepen the former, this criterion could be of use as well. We will give a more precise definition of this notion in Section 6.6.2.

6.5.2 Selection Procedures for New Types of A-Sets

On the basis of the newly introduced criteria, we are now able to express selection procedures that correspond to certain types of epistemic evaluation of scientific theories, that is, views of scientists participating in scientific debates.

Procedure 1. The underlying idea of this procedure is to select the argumentative core of the most explanatory scientific views or theories together with arguments that are used for attacking their rivals. It consists of the following steps:

[16] This is still, of course, a very rough account of the degree of being defended. It would get more realistic if we took into consideration a weighting of the attacks, since some attacks may be considered as more severe than others (see also our Discussion in Section 6.7.3).

1. Select all the conflict-free A-sets.
2. Out of these, select the most explanatory A-sets.
3. Out of these, select the most defended A-sets.
4. Out of these, select the maximal A-sets (w.r.t. set-inclusion).

Applied to our example, the procedure delivers the A-set $\{b, f, g, h, i, l, k\}$. First, by selecting conflict-free A-sets, we make sure that we distinguish between the rivaling theories. Second, we choose from these the most explanatory powerful ones. Third, by choosing the most defended ones of these, we select the least problematic of the explanatory powerful theories. Finally, by choosing the maximal ones from the latter selection, we make sure we include as many mutually compatible arguments as possible, thus also including those which are used to attack the rivaling theories.

Procedure 2. The idea underlying this procedure is to select the explanatory core of the most explanatory theories.

1. Select all the conflict-free A-sets.
2. Out of these, select the most explanatory A-sets.
3. Out of these, select the most defended A-sets.
4. Out of these, select the explanatory deepest A-sets.
5. Out of these, select the minimal A-sets (w.r.t. set-inclusion).

In the first three steps we proceed analogous to Procedure 1. By selecting the explanatory deepest A-sets we want to make sure we include the explanatory gist of the given theory. Finally, by choosing the minimal of these, we preserve only those arguments belonging to such an explanatory core, while disregarding, for example, the arguments used only for attacking the rivaling theories, i.e. arguments that don't have an explanatory or defensive function. Applied to our example this procedure delivers the A-set $\{f, g, h, i\}$ (as the reader can easily verify).

Even though both of our procedures prioritize the criterion of explanatory power to the criterion of defense, in some contexts we may wish to reverse this order, and even use the full defense criterion. For instance, when evaluating which theories should be accepted (and not only pursued), we may want to allow only for theories that can be fully defended. Such a procedure would begin with the selection of defended A-sets, followed by the selection of the most explanatory ones of these. Obviously, by different combinations of the criteria we may obtain different procedures suitable for different epistemic contexts.

Let us also remark that the procedures offered above are of a sequential or vertical nature: selections with respect to various criteria are applied step-wise. It is also possible to select "horizontally" by making use of weighting functions for A-sets. Let us give a simple example. Given an EAF $\langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow, \sim \rangle$ and a conflict-free A-set A we define:

$$\pi(A) = \mu_d \frac{|\mathcal{A} \setminus \alpha(A)|}{|\mathcal{A}|} + \mu_e \frac{|\epsilon(A)|}{|\mathcal{X}|},$$

where $\alpha(A)$ is the set of attackers of A against which it cannot defend itself. Moreover, μ_d and μ_e are numerical weights that model the importance we attach to the criteria defendedness and explanatory power respectively. Let us apply π to some A-sets from Example 6.5.1 where $\mu_d = \mu_e = 1$: for the Drift we have e.g. $\pi(\{b, f, g, h, i, k, l\}) = \pi(\{i, f, g, h\}) = \frac{10}{11} + \frac{3}{3}$, while for contractionism we have $\pi(\{a, c, d, j\}) = \pi(\{a, c, d\}) = \frac{10}{11} + \frac{2}{3}$. Of course, horizontal and vertical selection mechanisms may be combined. For instance, we could first select A-sets that maximize π and then select the maximal ones out of these. In the example above we would end up with $\{b, f, g, h, i, k, l\}$.

In the following section we will give a more precise formal representation of the comparisons in view of explanatory power and explanatory depth, which will also allow for a refinement of different aspects of these two procedures.

6.6 A Formal Account of Explanatory Properties

6.6.1 Explanatory Power

Let us now properly define the two ways of comparing the explanatory power of A-sets that have been introduced in the previous section and point out possible refinements for each of them.

Definition 6.6.1. *Comparing the explanatory power in a qualitative sense:* A is explanatory stronger than A', in signs $A' \sqsubset_e A$, iff the set of explananda for which A offers an explanation is a super-set of the set of explananda for which A' offers an explanation: $\epsilon(A') \subset \epsilon(A)$. This notion was used in Section 6.5.

Definition 6.6.2. *Comparing the explanatory power in a quantitative sense:* A is explanatory stronger than A', in signs $A' \sqsubset_c A$, iff A explains numerically more explananda than A': $|\epsilon(A')| < |\epsilon(A)|$.[17]

Example 6.6.1. Let $\langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow, \sim \rangle$ be an EAF where $\mathcal{X} = \{e_1, \ldots, e_{10}\}$. Let A and A' be preferred A-sets for which $\epsilon(A) = \{e_1, \ldots e_8\}$ and $\epsilon(A') = \{e_6, \ldots, e_{10}\}$. Note that $A' \sqsubset_c A$ since A explains 8 out of 10 explananda but A' only explains 5. However, $A' \not\sqsubset_e A$ since $\epsilon(A') \not\subset \epsilon(A)$.

Sometimes the comparative measures of explanatory power offered in Definitions 6.6.1 and 6.6.2 are too strict. To see this suppose that there is a third preferred A-set A'' in our Example 6.6.1 for which $\epsilon(A'') = \{e_4, \ldots, e_{10}\}$. Note that $A'' \sqsubset_c A$. However, A numerically explains only one explanandum more than A''. Often if the explanatory power of two theories is not very different this is not a sufficient reason for preferring one over the other.

For a more refined approach to representing both comparative notions of explanatory power, we generalize them by introducing a threshold value τ. For the quantitative notion we define $A \sqsubset_c^\tau A'$ iff $|\epsilon(A')| - |\epsilon(A)| > \tau$ where τ is a constant. Note that $\sqsubset_c^0$ is equivalent to $\sqsubset_c$. By introducing a threshold value

[17] It is easy to see that $\sqsubset_c$ and $\sqsubset_e$ are strict preorders on $\wp(\mathcal{A}) \times \wp(\mathcal{A})$. (A strict preorder is a irreflexive and transitive binary relation.) Obviously $A \sqsubset_e A'$ implies $A \sqsubset_c A'$.

τ, for instance 1, both A-sets A and A'' are equally explanatory strong with respect to $\sqsubset_c^1$. This introduces an interesting option for the modeling of states in science where different scientific groups are characterized by similarly strong, but nevertheless incompatible explanatory features. It is easy to see that $\sqsubset_e$ can be generalized in a similar way using threshold values.

Example 6.6.2. Let us return to our Example 6.5.1. Take for instance the two conflict-free A-sets: $A_7 = \{a, c, d\}$ and $A_8 = \{b, f, g, h, i\}$. It is easy to see that A_8 has a greater explanatory power than A_7 since it offers an explanation of e_1, e_2, and e_3, while A_7 only offers an explanation of e_1 and e_2. Formally speaking, $A_7 \sqsubset_e A_8$ as well as $A_7 \sqsubset_c A_8$.

Nevertheless, the explanatory power of the two A-sets in question is similar (although in our simplified modeling $\epsilon(A_7) \subset \epsilon(A_8)$): A_8 explains only one explanandum more than A_7. By introducing a threshold $\tau = 1$, we obtain both extensions – A_7 and A_8 – as maximal elements of $\sqsubset_c^1$ (with respect to all conflict-free A-sets): due to their high and similar explanatory power they are both acceptable according to this notion. Such a rendering would correspond, for instance, to the view that geology in the first half of the twentieth century was in a multi-paradigm state where both contractionism and the Drift (as well as permanentism) were mutually rivaling paradigms (see e.g. (Stewart 1990, p. 139)), in contrast to the above, strict rendering which corresponds to the preference on the Drift as a more explanatory powerful conception. This is due to the fact that the explanatory power of the two camps was very similar such that, given an appropriate threshold τ, also for a more complete and realistic modeling of this debate, A-sets representing both views would be selected.

6.6.2 Explanatory Depth

As we have seen in the previous section, we can also compare A-sets with respect to their explanatory depth. This is important, for instance, in cases in which we have two A-sets with the same explanatory power, but one of which offers a deeper explanation of some explananda than the other one. The formal definition of explanatory depth is as follows:

Definition 6.6.3. Given two A-sets A_1 and A_2, we say that A_2 is *explanatory at least as deep as* A_1, in signs $A_1 \sqsubseteq_d A_2$ iff for every explanation $X_1[e]$ of $e \in \epsilon(A_1)$ offered by A_1 there is an explanation $X_2[e]$ offered by A_2 such that $X_1 \prec X_2$ or $X_1 = X_2$. We say that A_2 is *explanatory deeper* than A_1, written $A_1 \sqsubset_d A_2$, iff $A_1 \sqsubseteq_d A_2$ but it is not the case that $A_2 \sqsubseteq_d A_1$.

For instance, in Example 6.5.1, $\{f, g, h, i\}$ is explanatory deeper than $\{f, g, h\}$.

In addition to their application in the selection procedures mentioned in the previous section, our criteria for explanatory power and explanatory depth may be interesting in capturing some philosophical notions as well. For example, Thagard's concepts of *broadening* – explaining new facts, and *deepening* – explaining why the theory works (Thagard 2007, p. 29) correspond to our notions

of explanatory power and explanatory depth. Similarly, we can account for the notions of "horizontal" and "vertical explanations" (see (Bermúdez 2005)) by representing them, respectively, with our notions of primary explanation and its deepening.[18]

6.7 Discussion

In this section we address questions that are relevant for situating our framework in the broader context of philosophy of science and methodology, Dung's abstract argumentation, as well as in the context of other formal accounts of explanatory reasoning. Therefore we shall clarify the relevance and the novelties of EAFs, as well as some possible enhancements.

6.7.1 EAFs and the Argumentative Shift in Methodology

Discussions in the field of philosophy of science and scientific methodology in the last couple of decades have witnessed a growing conviction that a rule-based algorithmic approach to theory appraisal is problematic. One possible attempt to preserve the normative idea of rationality in spite of abandoning the idea of a static, universally applicable scientific method can be found in more rhetorically minded approaches to scientific reasoning, such as Pera's (1994) or Dascal's (2000). Instead of an algorithmic assessment of scientific theories, Pera and Dascal emphasize the evaluation in view of the argumentative context underlying the given episode in the history of science. Similarly, Longino points out that a "method must [...] be understood as a collection of social, rather than individual, processes, so the issue is the extent to which a scientific community maintains critical dialogue" (Longino 1990, p. 76). While formal approaches to scientific reasoning have been mainly focused on the logical form of arguments (that is, the nature of the inference relation), both Pera and Dascal show that scientific debates (Dascal's *controversies*) are typically not resolved by the derivational reasoning that is characteristic for logic but rather by scientists exchanging arguments and trying to convince each other by giving reasons that substantiate their points:

> The contenders pile up arguments they believe increase the weight of their positions vis a vis the adversaries' objections, thereby leading, if not to deciding the matter in question, at least to tilting the 'balance of reason' in their favor. Controversies are neither 'solved'

[18] According to Bermúdez, a "horizontal explanation is the explanation of a particular event or state in terms of distinct (and usually temporarily antecedent) events or states." (Bermúdez 2005, p. 32). For example, a horizontal explanation providing an answer to the question why the window broke when it did, might call upon the baseball's hitting it and a generalization about windows tending to break when hit by a baseball. However, if we ask why the mentioned generalization holds, that is, what features of the physical structure of glass make it fragile in such circumstances – we are asking for an explanation of the grounds of the given horizontal explanation. Such explanations Bermúdez calls vertical explanations (ibid, p. 32-33).

> nor 'dissolved'; they are resolved. Their resolution may consist in the acknowledgment (by the contenders or by their community of reference) *that enough weight has been accumulated in favor of one of their positions, or in the emergence (thanks to the controversy) of modified positions acceptable to the contenders, or simply in the mutual clarification of the nature of the differences at stake.* (Dascal 2000, p. 165, italics added)

Our account of EAFs is supposed to mirror the idea underlying such an argumentative approach to scientific controversies in a formal way. Various possible enhancements (which will be mentioned in subsection 6.7.3) allow for a framework that reflects such a rhetorically minded approach to scientific debates in a more refined way. However, in contrast to an informal analysis such as Pera's *Dialectics*, which deals with rhetorical aspects of arguments in scientific debates in terms of classifying argument types and explicating their roles (see also Chapter 6), our approach abstracts from the concrete type of arguments by focusing only on their roles as being attacks or explanations. This allows us to inhabit a formal middle-ground for the modeling of the "tilting of the 'balance of reason'" by means of selection procedures defined with the help of our framework. Thus, EAFs (and more generally speaking, abstract argumentation) can be considered to complement the informal theories of argumentation by representing a formal tool that can serve to rationally reconstruct scientific debates from an argumentative point of view.

6.7.2 The Novelty of EAFs

What are the main novelties of our framework? This question should be answered in view of the research done in abstract argumentation frameworks, as well as in view of other formal accounts of explanatory reasoning.

With regard to the former, it could be argued that since our explanatory arrow is a kind of support relation, systems such as the bipolar one (see (Cayrol and Lagasquie-Schiex 2005)), which also feature an argumentative support relation, might be sufficient to model the notions introduced by our framework (such as explanations, explanatory power, explanatory depth, etc.). Nevertheless, the presence of the set of explananda $\mathcal{X}$ in EAFs makes an important difference. Placing explananda outside of the set of arguments makes it possible not only to express notions such as the explanatory power of an A-set, but also to represent alternative explanations of the same phenomenon and to compare the explanatory virtues of different A-sets. Moreover, on the basis of such an enhancement we are able to formulate new selection types in view of explanatory properties of the arguments, which are more suitable for the evaluation of scientific views than the standard Dung's selection types. Thus, our explanatory relation cannot be substituted by the already existing support relation.

With regard to other formal accounts of explanatory reasoning, it is important to notice that there are different levels of abstraction on which formal

representations can be based. First of all, we can formally analyze explanatory reasoning by focusing on the nature of the inferential relation present in explanations. This will give us, for instance, logical systems of abduction (see e.g. (Aliseda 2006)). Next, we can obtain formal representations by abstracting away from the logical properties of explanatory reasoning and focusing on the explanatory coherence of the propositions constituting a certain cognitive system. Thagard's account of explanatory coherence (Thagard 1992) and its implementation in the computer program ECHO is an example of such an approach.[19] Finally, if we abstract away from the propositional level, we can represent explanatory reasoning in terms of arguments taken in an abstract sense of the term, that is, without analyzing the specific type of inferential relations involved in them. This kind of approach is the one employed in EAFs. An important merit of such an approach is that it allows for a transparent representation of scientific debates, while at the same time offering handy tools for evaluating A-sets in view of their argumentative properties (e.g. being conflict-free and defended) and explanatory virtues. Moreover, by abstracting away from the specific nature of inferential relations in argumentative reasoning, we are able to model debates in which not every argument is necessarily based on a valid inference (and which may lead to it being attacked by a counterargument). Hence, by applying such an approach to scientific reasoning, we have not only introduced a novelty in the field of abstract argumentation, but also in the field of formal representations of scientific reasoning, which has, to our knowledge, so far not been linked with argumentation in an abstract sense of the term.[20]

6.7.3 Enhancing EAFs

In order to allow for an even more realistic modeling of scientific debates, our framework can be enhanced in different ways. For example, we can easily introduce a support relation (presented in (Cayrol and Lagasquie-Schiex 2005))

[19] It is important to notice that even though both Thagard's explanatory coherence and our EAFs aim at modeling the comparison of cognitive systems in terms of their explanatory virtues, there is a number of differences between these two accounts. For example, while the basic unit of Thagard's coherentism is a proposition, our framework is constituted of arguments and explananda; while Thagard's notion of acceptability is defined in terms of explanatory coherence, we speak of different types of acceptability, defined in terms of defensibility and certain explanatory properties; while Thagard does not model the dynamics of argumentation and thus cannot model the idea of an argumentative defense, we can; while the distinction of the evaluation of theories in the context of pursuit and in the context of acceptance is not explicated in his account, we have shown that such a distinction can be made in EAFs; finally, our graphical representation is quite different from Thagard's and may be considered as more transparent when it comes to the representation of scientific debates.

[20] Even though in this thesis we have been primarily concerned with scientific explanations, philosophy of science is not the only novel domain in which abstract argumentation in terms of EAFs may be fruitfully applied. Another interesting application field is expert systems. For example, Moulin et al. (Moulin et al. 2002) argue that in order to justify and "convince their users of the validity of their recommendations ..., artificial agents should also be equipped with explanation and argumentation capabilities." (p. 172).

which runs over the set of arguments.[21] Another interesting enhancement would be to introduce a weighting on arguments (by means of values (Bench-Capon 2002), (Bench-Capon 2003), preferences (Amgoud and Cayrol 1998), or audiences (Bench-Capon et al. 2007)), evidential support (Oren and Norman 2008), or to introduce different types of joint attacks (Nielsen and Parsons 2006) or fuzziness (Janssen et al. 2008) that allows, for instance, for the relaxation of the standards of being conflict-free and of being defended. Introducing nested attacks in a hierarchical manner into EAFs (see (Modgil 2006, 2009)) would allow for a formal distinction and an analysis of the interplay between arguments given on the object level and methodological arguments addressing the former ones. Moreover, our explanatory relation could be refined by formally distinguishing between two types of deepening explicated in Section 6.4.1. Similarly to the above mentioned nested attacks, we could allow for a nested explanatory relation: $a \dashrightarrow b$ would in that case indicate that a explains one of the premises of b, while $a \dashrightarrow (b \dashrightarrow c)$ would indicate that a explains the explanatory link between b and c.

Another possible enhancement is related to the property of our framework that explananda are not supposed to be a matter of dispute. Even though this is often so, there are rare cases in which not all involved parties agree about what is to count as a *significant* or even *valid* explanandum in the given field. For example, some geologists criticized Wegener for pointing out that his theory of continental drift explained the jigsaw-fit of continental coastlines, since such a match was, according to them, not at all obvious and did not represent a significant explanandum for geological sciences. In order to model such disputes, EAFs could be enhanced by allowing preferences or values on explananda, so that different opinions of scientists can be represented in that way. However, if we want to model cases in which explananda can be rejected as invalid on an argumentative basis, that is, if we want to allow for a dispute on explananda we could enhance EAFs by allowing attack relation to run not only over arguments, but also to go from arguments to explananda (that is, $\rightarrow \subseteq \mathcal{A} \times (\mathcal{A} \cup \mathcal{X})$). As a result, explananda could be both criticized and defended, while the modeling of explanatory virtues would have to be adjusted in order to account for the fact that different sets of explananda are acceptable with respect to different sets of arguments.

As we have mentioned in the introduction, an important virtue of abstract argumentation is that a basic framework can easily be enhanced in various ways. EAFs as presented in this chapter clearly provide an idealized modeling with regard to the subtleties that occur in real scientific debates. By listing possible enhancements of EAFs we have suggested in which way a more realistic modeling could be obtained.

[21] After all, there are types of argumentative supports that are not explanatory in nature and can hence not be represented by means of our explanatory relation.

6.8 Conclusion

In this chapter we have presented the Explanatory Argumentation Framework (EAF), obtained by enhancing Dung's Abstract Argumentation Framework with explanatory capabilities. We have motivated such an enhancement by pointing out, on the one hand, the close relation between argumentation and explanation, and on the other hand, the usefulness of our framework in the modeling of scientific debates. We have demonstrated that EAFs allow for a dynamic view on explanatory reasoning by involving both justificatory and explanatory arguments. The relation to scientific debates has been explicated by showing how multiple different scientific views can be modeled by making use of different criteria for selecting arguments. In this way EAFs are able (i) to model the criticism inherent to scientific debates in terms of counter-arguments, (ii) to model alternative competing explanations, and (iii) to evaluate and compare the explanatory features offered by the competing scientific views. Moreover, different selection procedures that can model different epistemic and methodological preferences regarding theory choice can be formulated in our framework.

Chapter

7 Adaptive Logic Framework for Abstract Argumentation

✎ *This chapter is based on a paper published under the name "Towards the Proof-Theoretic Unification of Dung's Argumentation Framework: An Adaptive Logic Approach" in the Journal of Logic and Computation (Straßer and Šešelja 2011). The paper is co-authored by the first author Christian Straßer. We are indebted to Diderik Batens, Joke Meheus, and two anonymous reviewers for valuable comments on the former draft of this paper.*

Summary This chapter presents a unifying adaptive logic framework for abstract argumentation. It consists of a core system for abstract argumentation and various adaptive logics based on it. These logics represent in an accurate sense all standard extensions defined within Dungs abstract argumentation system with respect to sceptical and credulous acceptance. The models of our logics correspond exactly to specific extensions of given argument systems. Additionally, the dynamics of adaptive proofs mirror the argumentative reasoning of a rational agent. In particular, the presented logics allow for external dynamics, i.e. they are able to deal with the arrival of new arguments and are therefore apt to model open-ended argumentations by providing provisional conclusions.

7.1 Introduction

In the previous chapter we have presented the abstract argumentation frameworks (AFs) as fruitful for modeling scientific reasoning. But which type of reasoning governs the selections of arguments in AFs? One way to answer this question is by developing a logical framework for such a reasoning.

Dung himself remarked that "Logic-based knowledge bases can be viewed as argumentation systems where the knowledge base is coded in the structure of the arguments and *the logic* is used to determine the acceptability of the

arguments." (Dung 1995, p. 856, italics added). Moreover, according to John L. Pollock: "Constructing arguments is one thing. Deciding which conclusions to accept is another. ... The conclusions that ought to be believed are those that are undefeated." (Pollock 1987, p. 7). Hence, in a logical system modeling the reasoning that underlies argumentation procedures, the focus is not on the derivation of propositions from other propositions, but on the derivation of arguments, and thus of the conclusions made by them, that are considered acceptable in view of the whole set of arguments constituting the given debate.

In this chapter we offer a proof-theoretic framework for all the standard selections of arguments (that we have in the previous chapter called A-sets) in AFs. Since this chapter will be more technical in character, we will introduce AFs once again, and by using the terminology that is standard in the literature on AFs, coming mainly from the research in artificial intelligence and computer science. For instance, A-sets will now be called extensions of a particular extension type.[1] Let us first recall the main features of AFs.

The key notion of Dung's account is the acceptability of arguments. Dependent on the criteria for acceptance, it is possible to formulate different semantics. These define a number of extensions representing sets of acceptable arguments, such as *admissible*, *grounded*, *complete*, *(semi)-stable* and *preferred* extensions. Dung's system has been extended and generalized in various respects. We name just a few: preferences (Amgoud and Cayrol 1998), values (Bench-Capon 2002) and audiences (Bench-Capon et al. 2007) in the sense of Perelman (Perelman and Olbrechts-Tyteca 1969) have been introduced, joint attacks have been enabled (Nielsen and Parsons 2006), the system has been used for an improved account to default reasoning (Bondarenko et al. 1997, Dung and Son 1996), it has been applied to multi-agent systems (Coste-Marquis et al. 2007, Bench-Capon 2003), new semantics/extensions have been presented (Dung et al. 2007, Caminada 2006), game-theoretic approaches have been developed (Dunne and Bench-Capon 2003), etc.

This chapter offers an adaptive logic framework with a specific core axiomatic system, on the basis of which we define logics for obtaining all the standard extension types of Dung's account with respect to the skeptical and the credulous acceptability of arguments. Adaptive logics were initially developed by Diderik Batens. Their first application was the handling of inconsistent premises. Since then they have been actively researched and used in order to explicate a variety of defeasible reasoning forms (for a recent survey see (Batens et al. 2009), for an explication of their meta-theory see (Batens 2007)). The main idea of adaptive logics (ALs) is to interpret a premise set "as normally as possible", given a certain standard of normality. Based on the so-called lower limit logic (**LLL**), they select certain **LLL**-models of a given premise set which satisfy the standard of normality. Syntactically they enrich the derivative power of the **LLL** by allowing for certain rules to be applied conditionally.

[1] The logics for EAFs could be formulated in a similar way as those that will be presented in this chapter. Hence, the work presented here can be considered as the basis for formulating different types of logics for AFs and their enhancements.

In case a condition turns out to be unsafe, formulas derived on this very condition are marked in the proof and thus not considered as being derived anymore. Markings in ALs come and go while we reason on. Indeed, adaptive proofs are dynamic in two ways: internally in the sense that during the reasoning process certain conditions might turn out to be unsafe/safe (again) while we get more insight into the given premises, externally in the sense that the introduction of new information in the form of new premises may alter our treatment of certain conditions and thus alter the markings in the proof. In argumentation the situation is similar: as rational debaters we introduce an argument a in such a way that we are willing to withdraw it under certain circumstances. For instance in the case that it is conflicting with other arguments, or in the case that we cannot defend it against certain counterarguments. However, at a later point, a new argument which defends the attacked a, might enter the scene, and cause the acceptance of the latter one again. Therefore, argumentation is in a similar way dynamic as adaptive logic proofs: internally in the sense that the progressing analyses of the relationship of given arguments might alter our choice for accepted arguments, and externally in the sense that the introduction of new arguments might make us reconsider the acceptance of some arguments.

As we have already mentioned, Dung defined various extension types which select certain subsets of (non-conflicting) arguments with respect to given criteria. This mirrors the semantic selection of adaptive logics: while the **LLL** defines minimal criteria that have to be fulfilled in every model (such as the absence of conflicts between the validated arguments and the property that every validated argument is defended by the other validated arguments against all possible attacks), the adaptive enhancements refine the semantic selection by modeling the criteria given by the various extension types. Furthermore, the dynamic adaptive proofs model the reasoning process leading to these selections. In summary, ALs are very suitable for providing a unifying logical framework for abstract argumentation. Thus, the logics for abstract argumentation which will be presented in this chapter are adaptive logics. They employ the core set of axioms as **LLL** and define different standards of normality. In this way we can obtain logics for *admissible, complete, preferred, (semi-)stable* as well as *grounded* extensions.

One of the main advantages of our approach compared to other proposals for proof theories for abstract argumentation (see (Vreeswijk and Prakken 2000, Cayrol et al. 2003, Dung et al. 2002)) lies in its unifying power. A single framework is able to capture all standard semantics/extensions with respect to both skeptical and credulous acceptance, which makes it an ideal logical surrounding for their comparison, further elaboration, enhancements and generalizations.

In addition, our system represents a contribution to the research done in applications of adaptive logics to different dialogical contexts (for a survey see (Batens 2001)). Furthermore, this chapter confirms the claim that the adaptive logic program offers a general and unifying framework for nonmonotonic and defeasible logics (as has recently been argued for in (Batens et al. 2009)).

The chapter is structured in the following way. First we present Dung's ab-

stract argumentation framework in Section 7.2. In Section 7.3 we introduce the reader step-wise into our adaptive logic framework for abstract argumentation. We begin by presenting a formal language and the basic axiomatic system in Subsection 7.3.1. Furthermore, we discuss how abstract argumentation frameworks can be represented as premise sets in Subsection 7.3.2. Subsection 7.3.3 contains representational requirements for our logics. In the remainder of Section 7.3 we introduce the reader paradigmatically, on the basis of preferred extensions and by means of examples, into the terminology and the modus operandi of adaptive logics. In Section 7.4 we define all the adaptive logics for the various extension types with respect to skeptical acceptance and state the corresponding representational results. Section 7.5 features all the logics for credulous acceptance and the respective representational results. In Section 7.6 we localize the adaptive logic approach within the field of logical representations of abstract argumentation and point out some advantages. All the meta-proofs for our results are presented in (Straßer and Šešelja 2009).

7.2 Dung's Argumentation Framework - Key Terms

We will use lower case letters $a_1, a_2, a_3, \ldots$ for arguments and lower case fraktur letters $\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \ldots$ as meta-variables for arguments. Let $\mathcal{A}_n =_{\text{df}} \{a_1, a_2, \ldots, a_n\}$. In (Dung 1995) Dung defined his abstract argumentation frameworks as follows:[2]

Definition 7.2.1. A *finite argumentation framework (AF)* is a pair $\langle \mathcal{A}, \rightarrow \rangle$ where $\mathcal{A} \subseteq \mathcal{A}_n$ is a finite set of arguments, and $\rightarrow \subseteq \mathcal{A} \times \mathcal{A}$ is a relation between arguments. The expression $\mathfrak{a} \rightarrow \mathfrak{b}$ is pronounced as "$\mathfrak{a}$ attacks $\mathfrak{b}$".

Given an AF $\langle \mathcal{A}, \rightarrow \rangle$ we are particularly interested in giving an account of reasonable choices of arguments in $\mathcal{A}$: a minimal criterion is, for instance, that no argument in a selection S should attack another argument in S. Of course, more interesting selection types can be defined:

Definition 7.2.2. Given an argumentation framework $\mathsf{A} = \langle \mathcal{A}, \rightarrow \rangle$ we define the following notions.

(i) An argument $\mathfrak{a}$ is *attacked* by a set of arguments $B \subseteq \mathcal{A}$ iff there is a $\mathfrak{b} \in B$ such that $\mathfrak{b} \rightarrow \mathfrak{a}$.

(ii) An argument $\mathfrak{a}$ is *acceptable* with respect to a set of arguments $C \subseteq \mathcal{A}$, iff every attacker of $\mathfrak{a}$ is attacked by C. It is said that C *defends* $\mathfrak{a}$.

(iii) A set of arguments $S \subseteq \mathcal{A}$ is *conflict-free* iff S doesn't attack any argument in S.

(iv) A conflict-free set of arguments $S \subseteq \mathcal{A}$ is *admissible* iff each argument in S is acceptable with respect to S.

[2] We restrict the discussion in this chapter to the finite case, i.e., to argumentation frameworks with a finite number of propositional letters.

(v) A set of arguments $S \subseteq \mathcal{A}$ is a *preferred extension* iff it is a maximal (w.r.t. $\subseteq$) admissible set.

(vi) A conflict-free set of arguments $S \subseteq \mathcal{A}$ is a *stable extension* iff it attacks every argument in $\mathcal{A} \setminus S$.

(vii) An admissible set of arguments $S \subseteq \mathcal{A}$ is a *complete extension* iff $F(S) = S$, where $F(S) =_{\text{df}} \{\mathfrak{c} \mid S \text{ defends } \{\mathfrak{c}\}\}$.

(viii) A set of arguments $S \subseteq \mathcal{A}$ is a *grounded extension* iff it is the minimal (w.r.t. $\subseteq$) complete extension.

(ix) A complete extension $S \subseteq \mathcal{A}$ is a *semi-stable extension*[3] iff $S \cup S^+$ is maximal (w.r.t. $\subseteq$), where S^+ is the set of all arguments in $\mathcal{A} \setminus S$ which are attacked by S.

(x) A set of arguments $S \subseteq \mathcal{A}$ is *credulously accepted* according to preferred [(semi)-stable, complete or grounded] semantics (w.r.t. A) iff it is contained in at least one preferred [(semi)-stable, complete or grounded] extension of A.

(xi) A set of arguments $S \subseteq \mathcal{A}$ is *skeptically accepted* according to preferred [(semi)-stable, complete or grounded] semantics (w.r.t. A) iff it is contained in every preferred [(semi)-stable, complete or grounded] extension of A.

Suppose we select a conflict-free set $E \subseteq \mathcal{A}$. There are two types of arguments in $\mathcal{A} \setminus E$ which are not selected. On the one hand, arguments in E^+ which are attacked by the selected arguments and on the other hand the ones that are not attacked by E, i.e., arguments in $\mathcal{A} \setminus (E \cup E^+)$. We call the former arguments *defeated*, since they are attacked by at least some of our selected arguments E. Admissibility requires that the set of defeated arguments for a given selection of arguments S consists at least of all the attackers of S. Opposite to attacks, we only speak of a defeat in view of a given selection of arguments: $\mathfrak{a}$ attacks $\mathfrak{b}$ iff $\mathfrak{a} \rightarrow \mathfrak{b}$, while $\mathfrak{b}$ is defeated iff there is a selected argument $\mathfrak{a}$ that attacks $\mathfrak{b}$. It would be misleading to confuse attack and defeat: an agent may (argumentatively) attack another agent, but we only consider the attack as a defeat if the argument used for the attack is considered as valid.[4]

[3] Semi-stable semantics were defined in (Caminada 2006) and are equivalent to Verhijs' *admissible stage extensions* in (Verheij 1996).

[4] Our notion of defeat differs from the way defeat is defined in various preference or value based enhancements of Dung's abstract argumentation framework. Defeat is there usually defined as a binary relation between arguments which is a subset of the attack relation: a_1 defeats a_2 iff a_1 attacks a_2 and a_2 is not 'preferable' to a_1. The preferability of one argument over another is modeled in different ways: in terms of a preference relation between arguments in (Amgoud and Cayrol 2002), by allowing for arguments to attack an attack in (Modgil 2009), or in terms of mapping arguments into partially ordered values in (Bench-Capon 2003).

Figure 7.1 illustrates some basic relationships between the various extension types. For a more thorough study of them we refer the reader to the rich literature mentioned in the introduction.

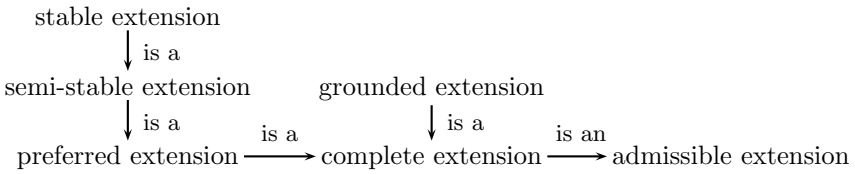


Figure 7.1: The relationship between the extensions types.

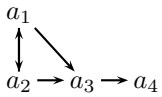


Figure 7.2: An attack-diagram.

Argumentation frameworks $\langle \mathcal{A}, \rightarrow \rangle$ are often represented by directed graphs, the so-called attack-diagrams (see Figure 7.2). The nodes are arguments in $\mathcal{A}$ and there is an edge from $\mathfrak{a}$ to $\mathfrak{b}$ iff $(\mathfrak{a}, \mathfrak{b}) \in \rightarrow$.

Example 7.2.1. We will demonstrate the concepts just introduced with the attack-diagram in Figure 7.2. The following table lists the extensions belonging to the extension types introduced in Definition 7.2.2.

admissible	preferred	semi-stable	complete	grounded
$\varnothing, \{a_1\}, \{a_2\}, \{a_1, a_4\}, \{a_2, a_4\}$	$\{a_1, a_4\}, \{a_2, a_4\}$	$\{a_1, a_4\}, \{a_2, a_4\}$	$\varnothing, \{a_1, a_4\}, \{a_2, a_4\}$	$\varnothing$

Stable extensions are, unlike the other extension types, not guaranteed to exist. Semi-stable extensions (see (Caminada 2006)) improve on that: they are guaranteed to exist, and in case stable extensions exists the semi-stable extensions are identical to them.[5] Moreover, there is one unique grounded extension.

7.3 A Logic for Abstract Argumentation

In this chapter we are going to present propositional logics for abstract argumentation. We will present for each extension type $\mathcal{E}$ (such as admissible,

[5]The following fact offers an alternative definition of semi-stable extensions in terms of admissible sets S for which $S \cup S^+$ is maximal: Let $\mathsf{A} = \langle \mathcal{A}, \rightarrow \rangle$ be an AF and $S \subseteq \mathcal{A}$. S is a semi-stable extension iff S is an admissible set of arguments for which there is no admissible set of arguments $T \subseteq \mathcal{A}$ such that $T \cup T^+ \supset S \cup S^+$. The statement is proven in (Straßer and Šešelja 2009).

complete, preferred, etc.) a corresponding logic $\mathbf{L}_{\mathcal{E}}$. The major idea is that $\mathbf{L}_{\mathcal{E}}$ derives for a given argumentation framework $\mathsf{A} = \langle \mathcal{A}, \rightarrow \rangle$ all skeptically (resp. credulously) accepted arguments and that the models represent the extensions of type $\mathcal{E}$.

This section will make this idea more precise. It will introduce basic notions and the modus operandi of our logical framework for abstract argumentation. First, in Subsection 7.3.1, we will propose a formal language and the core set of rules for our logics. Subsection 7.3.2 offers a way to represent AFs as premise sets. We give a precise account of the representational requirements for our logics in Subsection 7.3.3. In the remainder of this section we will introduce the reader into the main ideas of adaptive logics by focusing on example cases, paradigmatically for preferred extensions. Section 7.4 will contain the definitions for all the other variants and the representational results.

7.3.1 Language and Rules

In order to represent a given AF A as a premise set we need a formal language which allows us to express the basic notions of abstract argumentation. The idea is, on the one hand, to represent arguments by propositional letters and, on the other hand, to enrich the language of classical propositional logic by a binary logical operator $\twoheadrightarrow$ where $\alpha \twoheadrightarrow \beta$ means that α attacks β. Since we represent arguments by propositional letters only, we restrict our language in such a way that only propositional letters are arguments of $\twoheadrightarrow$.[6] Formally the set of well-formed formulas $\mathcal{W}_n$ (where n is a natural number) is defined in the following way:

$$\begin{array}{lcl} \mathcal{V}_n & := & p_1 \mid p_2 \mid p_3 \mid \ldots \mid p_n \\ \mathcal{W}_n^{\twoheadrightarrow} & := & \langle \mathcal{V}_n \rangle \twoheadrightarrow \langle \mathcal{V}_n \rangle \mid \bot \twoheadrightarrow \langle \mathcal{V}_n \rangle \\ \mathcal{W}_n & := & \bot \mid \langle \mathcal{V}_n \rangle \mid \langle \mathcal{W}_n^{\twoheadrightarrow} \rangle \mid \neg \langle \mathcal{W}_n \rangle \mid \langle \mathcal{W}_n \rangle \wedge \langle \mathcal{W}_n \rangle \mid \langle \mathcal{W}_n \rangle \vee \langle \mathcal{W}_n \rangle \mid \langle \mathcal{W}_n \rangle \supset \langle \mathcal{W}_n \rangle \end{array}$$

$\mathcal{V}_n$ are the propositional letters of our language. We will in the remainder abbreviate $\neg(\alpha \twoheadrightarrow \beta)$ by $\alpha \nrightarrow \beta$.

Let us introduce the rules characterizing our core logic for abstract argumentation. First of all, it is obvious that if α is valid and it attacks β, $\alpha \twoheadrightarrow \beta$, then β should not be a consequence of our logic:

$$\frac{\alpha \quad \alpha \twoheadrightarrow \beta}{\neg\beta} \qquad (\mathrm{R}\twoheadrightarrow)$$

This rule guarantees the conflict-freeness of our consequences since by (R$\twoheadrightarrow$) we immediately get

$$\vdash (\alpha \twoheadrightarrow \beta) \supset (\neg\alpha \vee \neg\beta)$$

So, whenever $\alpha \twoheadrightarrow \beta$, either α or β is considered to be invalid. Moreover, in our language it is easy to express that an argument has been defeated by one

[6] We also allow for $\bot$ on the left hand side of $\twoheadrightarrow$. We will comment on this in a moment.

or more arguments. We define:

$$\mathsf{def}\,\beta =_{\mathrm{df}} \bigvee_{\alpha\in\mathcal{V}_n}\big(\alpha \wedge (\alpha \twoheadrightarrow \beta)\big) \tag{Def}$$

It is easy to verify that the following properties are immediate consequences of this definition and rule (R↠):

$$\vdash \mathsf{def}\,\alpha \supset \neg\alpha$$
$$\vdash \big(\alpha \wedge (\alpha \twoheadrightarrow \beta)\big) \supset \mathsf{def}\,\beta$$

The first property guarantees that, if an argument has been defeated, then it is supposed not to be validated by the logic. The second property assures that, if α and $\alpha \twoheadrightarrow \beta$ have been derived then β is defeated.

Remember that the idea behind admissible extensions is that a selected set of arguments S is required to defend itself. That is to say, in case an argument $\mathfrak{a}$ in S is attacked by another argument $\mathfrak{b}$, $\mathfrak{b} \to \mathfrak{a}$, then there is an argument $\mathfrak{c}$ in S which attacks $\mathfrak{b}$ or, in yet other words, $\mathfrak{b}$ is defeated. For our logics we can express this as follows:

$$\frac{\alpha \quad \beta \twoheadrightarrow \alpha}{\mathsf{def}\,\beta} \tag{Rad}$$

If we have α and $\beta \twoheadrightarrow \alpha$, then β is supposed to be defeated. Note that it would be insufficient to replace the conclusion $\mathsf{def}\,\beta$ of rule (Rad) by $\neg\beta$. In case an argument $\mathfrak{a}$ is selected and $\mathfrak{b}$ attacks $\mathfrak{a}$, $\mathfrak{b} \to \mathfrak{a}$, then it is not enough simply not to select $\mathfrak{b}$. By the requirement of admissible extensions the selected arguments have to defend themselves against all attackers. Thus, what is required in terms of our language is $\mathsf{def}\,\beta$. This ensures that there is an argument γ which attacks and thus defeats β. This is guaranteed by $\mathsf{def}\,\beta$ due to its definition $\bigvee_{\delta\in\mathcal{V}_n}(\delta \wedge (\delta \twoheadrightarrow \beta))$. The existence of a defeating argument γ of β would not be guaranteed were we to replace $\mathsf{def}\,\beta$ by $\neg\beta$ in the conclusion of (Rad).

An attentive reader might have noticed that our language also allows for $\bot \twoheadrightarrow \alpha$. This is helpful in order to express that a given propositional letter α corresponds to an argument in the given argumentation framework $\langle\mathcal{A}, \to\rangle$. The cardinality of $\mathcal{V}_n$ might be higher than the cardinality of $\mathcal{A}$ and thus there might be propositional letters which do not correspond to the given arguments $\mathcal{A}$. We express the fact that a propositional letter represents an argument by $\bot \twoheadrightarrow \alpha$. All propositional letters that do not represent an argument are guaranteed not to be validated by the following rule:

$$\frac{\bot \not\twoheadrightarrow \alpha}{\neg\alpha} \tag{R⊥}$$

Let us have a look at one last rule that will help us to represent complete extensions. The main idea behind this extension type is that any argument that is defended by a given selected set of arguments S is supposed to be in the selection S. That is to say, if S defends argument $\mathfrak{a}$, then $\mathfrak{a} \in S$. This can

be expressed by the following rule:

$$\frac{\bot \twoheadrightarrow \beta \quad \bigwedge_{\alpha\in\mathcal{V}_n}\big((\alpha \twoheadrightarrow \beta) \supset \mathsf{def}\,\alpha\big)}{\beta} \tag{RCo}$$

That β is defended by the set of validated arguments is expressed by $\bigwedge_{\alpha\in\mathcal{V}_n}((\alpha \twoheadrightarrow \beta) \supset \mathsf{def}\,\alpha)$: for every attacker α of β, α is supposed to be defeated. We also add the condition that β is actually representing an argument, $\bot \twoheadrightarrow \beta$, since, if it did not, then β would have no attackers and thus the other antecedent, $\bigwedge_{\alpha\in\mathcal{V}_n}(\alpha \twoheadrightarrow \beta \supset \mathsf{def}\,\alpha)$, would be valid. But, as already pointed out, we want to keep propositional letters which do not represent arguments out of the consequence sets of our logics.

The presented rules enable us to define the following two logics which will serve as **LLL**s for our adaptive systems.

Definition 7.3.1. $\mathbf{L_A}$ is classical propositional logic enriched by the rules (R$\twoheadrightarrow$), (Rad), and (R$\bot$). $\mathbf{L_C}$ is $\mathbf{L_A}$ enriched by (RCo).

We define the semantics for logics $\mathbf{L} \in \{\mathbf{L_A}, \mathbf{L_C}\}$ via an assignment function $v : \mathcal{V}_n \cup \mathcal{W}_n^{\twoheadrightarrow} \to \{0,1\}$ and an $\mathbf{L}$-valuation $v_M^{\mathbf{L}} : \mathcal{W}_n \to \{0,1\}$ determined by the assignment. We use an extended assignment function $v : \mathcal{V}_n \cup \mathcal{W}_n^{\twoheadrightarrow} \to \{0,1\}$ that assigns truth values to both, propositional letters and 'attacks', i.e., formulas in $\mathcal{W}_n^{\twoheadrightarrow}$. A model M is defined by an assignment function v. The following definitions are useful in order to define the $\mathbf{L_A}$-valuation based on v:

$$\begin{aligned}
v_{\mathrm{Rral}} &=_{\mathrm{df}} 1 - \max_{\alpha,\beta\in\mathcal{V}_n}\big(\min(v(\alpha), v(\alpha \twoheadrightarrow \beta), v(\beta))\big)\\
v_{\mathrm{Rbot}} &=_{\mathrm{df}} 1 - \max_{\alpha\in\mathcal{V}_n}\big(\min(v(\alpha), 1 - v(\bot \twoheadrightarrow \alpha))\big)\\
v_{\mathrm{Rad}} &=_{\mathrm{df}} 1 - \max_{\alpha,\beta\in\mathcal{V}_n}\big(\min\big(v(\alpha), v(\beta \twoheadrightarrow \alpha), 1 - \max_{\gamma\in\mathcal{V}_n}\big(\min(v(\gamma), v(\gamma \twoheadrightarrow \beta))\big)\big)\big),\\
v_{\mathsf{i}}^{\mathbf{L_A}} &=_{\mathrm{df}} \min(v_{\mathrm{Rral}}, v_{\mathrm{Rbot}}, v_{\mathrm{Rad}})
\end{aligned}$$

Note that v_{Rral} corresponds to our syntactical rule (R$\twoheadrightarrow$) in the sense that $v_{\mathrm{Rral}} = 1$ iff the assignment satisfies the semantic counterpart to (R$\twoheadrightarrow$). That is to say, $v_{\mathrm{Rral}} = 1$ iff v satisfies

$$\text{If } v(\alpha) = v(\alpha \twoheadrightarrow \beta) = 1, \text{ then } v(\beta) = 0. \tag{S$\twoheadrightarrow$}$$

The situation is analogous for v_{Rbot} and v_{Rad} with respect to the following properties:

$$\text{If } v(\alpha) = 1, \text{ then } v(\bot \twoheadrightarrow \alpha) = 1. \tag{S$\bot$}$$

$$\text{If } v(\alpha) = v(\beta \twoheadrightarrow \alpha) = 1, \text{ then there is a } \gamma \in \mathcal{V}_n \text{ for which } v(\gamma) = v(\gamma \twoheadrightarrow \beta) = 1. \tag{Sad}$$

We call an assignment $\mathbf{L_A}$-*intended* iff $v_{\mathsf{i}}^{\mathbf{L_A}} = 1$. In (Straßer and Šešelja 2009) it is shown that an assignment v is $\mathbf{L_A}$-intended iff v satisfies (S$\twoheadrightarrow$), (Sad) and

(S⊥).

We define the valuation function $v_M^{\mathbf{L}} : \mathcal{W}_n \to \{0,1\}$ paradigmatically for $\mathbf{L} = \mathbf{L_A}$. The one for $\mathbf{L_C}$ is defined in a similar way and can be found in (Straßer and Šešelja 2009). Where $\alpha, \beta \in \mathcal{V}_n$ and $\varphi, \varphi_1, \varphi_2 \in \mathcal{W}_n$ we define:

$$v_M^{\mathbf{L}}(\bot) = 0 \tag{s⊥}$$

$$v_M^{\mathbf{L}}(\alpha \twoheadrightarrow \beta) = v(\alpha \twoheadrightarrow \beta) \tag{s↠}$$

$$v_M^{\mathbf{L}}(\bot \twoheadrightarrow \alpha) = v(\bot \twoheadrightarrow \alpha) \tag{s⊥↠}$$

$$v_M^{\mathbf{L}}(\alpha) = \min(v_{\mathsf{i}}^{\mathbf{L_A}}, v(\alpha)) \tag{sPA}$$

$$v_M^{\mathbf{L}}(\varphi_1 \wedge \varphi_2) = \min(v_M^{\mathbf{L}}(\varphi_1), v_M^{\mathbf{L}}(\varphi_2)) \tag{s∧}$$

$$v_M^{\mathbf{L}}(\varphi_1 \vee \varphi_2) = \max(v_M^{\mathbf{L}}(\varphi_1), v_M^{\mathbf{L}}(\varphi_2)) \tag{s∨}$$

$$v_M^{\mathbf{L}}(\varphi_1 \supset \varphi_2) = \max(1 - v_M^{\mathbf{L}}(\varphi_1), v_M^{\mathbf{L}}(\varphi_2)) \tag{s⊃}$$

$$v_M^{\mathbf{L}}(\neg\varphi) = 1 - v_M^{\mathbf{L}}(\varphi) \tag{s¬}$$

Obviously, by (s↠) and (s⊥↠) the valuation inherits the truth values for 'attacks' in $\mathcal{W}_n^{\twoheadrightarrow}$ from the assignment function. Note that although (sPA) is of a rather complex form, it is fully determined by the assignment v. In the case $v_{\mathsf{i}}^{\mathbf{L_A}} = 1$, i.e., in the case that the assignment is $\mathbf{L_A}$-intended, the valuation takes over all truth values from the assignment for all formulas in $\mathcal{V}_n$. However, if $v_{\mathsf{i}}^{\mathbf{L_A}} = 0$, the valuation assigns to all propositional letters the truth value 0. Note that for a given AF A the empty selection is always an admissible extension. Thus, the valuation on the basis of a non-intended assignment corresponds to the empty extension. In (Straßer and Šešelja 2009) it is shown that $\mathbf{L_A}$-valuations satisfy (S↠), (Sad) and (S⊥).

Model validity and the semantic consequence relation are defined in the usual way. Where $\mathbf{L} \in \{\mathbf{L_A}, \mathbf{L_C}\}$, we define $M \vDash_{\mathbf{L}} \varphi$ iff $v_M^{\mathbf{L}}(\varphi) = 1$. We say a model M is an $\mathbf{L}$-model of $\Gamma \subset \mathcal{W}_n$ iff $M \vDash_{\mathbf{L}} \varphi$ for all $\varphi \in \Gamma$. We write $\mathcal{M}_{\mathbf{L}}(\Gamma)$ for the set of all $\mathbf{L}$-models of Γ. The semantic consequence relations $\Vdash_{\mathbf{L}}$ are defined in the usual way: $\Gamma \Vdash_{\mathbf{L}} \varphi$ iff for all $\mathbf{L}$-models M of Γ, $M \vDash_{\mathbf{L}} \varphi$.

Completeness and soundness for both logics, $\mathbf{L_A}$ and $\mathbf{L_C}$ are proven in (Straßer and Šešelja 2009).

7.3.2 Representing AFs as Premise Sets

Let us now see how to represent AFs in terms of premise sets. There is an easy and intuitive way to do so:

- First, we need to map the arguments of a given AF $\mathsf{A} = \langle \mathcal{A}, \rightarrow \rangle$, where $\mathcal{A} \subseteq \mathcal{A}_n$, into the set of propositional letters $\mathcal{V}_n$. Of course, we need at least as many propositional letters as we have arguments. A canonical way to do so is by $\lambda_n : \mathcal{A}_n \to \mathcal{V}_n, a_i \mapsto p_i$ for all $1 \leq i \leq n$. We can say that p_i represents, or corresponds to an argument a_i iff $a_i \in \mathcal{A}$.

- Second, we need to represent the attack relation. This can be simply done by adding to the premise set $p_i \twoheadrightarrow p_j$ iff $(a_i, a_j) \in \rightarrow$.

- It is furthermore important to indicate which propositional letters belong to the AF in question. We do this by adding $\bot \twoheadrightarrow p_i$ to the premise set iff $a_i \in \mathcal{A}$.

For premise sets constructed in this way we write Γ^n_{A}.

Example 7.3.1. For instance, the AF A from Example 7.2.1 is represented by the premise set $\Gamma^n_{\mathsf{A}} = \{p_1 \twoheadrightarrow p_2, p_2 \twoheadrightarrow p_1, p_1 \twoheadrightarrow p_3, p_2 \twoheadrightarrow p_3, p_3 \twoheadrightarrow p_4\} \cup \{\bot \twoheadrightarrow p_1, \bot \twoheadrightarrow p_2, \bot \twoheadrightarrow p_3, \bot \twoheadrightarrow p_4\}$ where $n \geq 4$.

For many applications it is interesting to choose a language that has more propositional letters than an argumentation framework $\langle \mathcal{A}, \rightarrow \rangle$ that is initially modeled. Some examples:

1. In order to model the argumentative reasoning of intelligent agents a system has to deal with more and more information in form of new arguments coming in. The initial setup is thus iteratively enriched as the argumentation proceeds. An argumentation can thus be seen as a sequence of argumentation frameworks $\mathsf{A}^1, \mathsf{A}^2, \ldots, \mathsf{A}^m$ where $\mathsf{A}^{i+1} = \langle \mathcal{A}^{i+1}, \rightarrow^{i+1} \rangle$ is an enhancement of $\mathsf{A}^i = \langle \mathcal{A}^i, \rightarrow^i \rangle$, i.e., $(\mathcal{A}^{i+1} \cup \rightarrow^{i+1}) \supset (\mathcal{A}^i \cup \rightarrow^i)$.

2. Abstract argumentation is a promising framework for applications such as machine learning (see (Možina et al. 2007)), belief revision (see (Falappa et al. 2009) for a survey), or decision theory (see (Amgoud and Vesic 2009)) since knowledge/belief bases may be represented by or with the help of argumentation frameworks.

For such applications it is obviously important to have enough propositional letters available in order to represent the successive stages $\mathsf{A}^1, \mathsf{A}^2, \ldots, \mathsf{A}^m$.

Furthermore, a logic has to be able to deal with new information arriving, resulting in the transition from A^i to A^{i+1}. That is to say, it has to allow for *external dynamics*. To simply apply a given algorithm producing the accepted argument in question again from scratch to A^{i+1} is cumbersome, especially since it doesn't model in any way the rationale of the rational agent going through this very transition. We will thus offer a dynamic proof procedure which by dynamic markings is able to model the reasoning of the agent in question throughout the sequence of updates she is exposed to. We will discuss this feature more in Subsection 7.3.6.

7.3.3 Representational Requirements

Given an extension type $\mathcal{E}$ (such as admissible, complete, preferred, etc.) and an AF A, what are our requirements for a logic for abstract argumentation $\mathbf{L}_{\mathcal{E}}$? What should its consequences look like, what should its models represent?

Let us presuppose that for the AFs $\langle \mathcal{A}, \rightarrow \rangle$ under consideration, $\mathcal{A} \subseteq \mathcal{A}_n$, and that $\mathbf{L}_\mathcal{E}$ is formulated in the language $\mathcal{W}_n$. This simply makes sure that we have enough propositional letters to logically represent the AFs in question.

We have two straightforward and intuitive representational requirements for a complete and sound logic $\mathbf{L}_\mathcal{E}$: a syntactical one and a semantical one. Let $\mathsf{A} = \langle \mathcal{A}, \rightarrow \rangle$ be an AF:

1. *Syntactic adequacy for skeptical (resp. credulous) acceptance*: We require that $\Gamma^n_\mathsf{A} \vdash_{\mathbf{L}_\mathcal{E}} p_i$ iff $a_i \in \mathcal{A}$ and a_i is skeptically (resp. credulously) accepted according to $\mathcal{E}$. Informally this simply means that a propositional letter is derivable iff it represents a skeptically (resp. credulously) acceptable argument (for $\mathcal{E}$).

2. *Semantic adequacy*: Let $\mathcal{M}_{\mathbf{L}_\mathcal{E}}(\Gamma^n_\mathsf{A})$ be the set of all $\mathbf{L}_\mathcal{E}$-models of Γ^n_A. We require that,

   a) For each $\mathcal{E}$-extension $E \subseteq \mathcal{A}$ of A there is a model $M \in \mathcal{M}_{\mathbf{L}_\mathcal{E}}(\Gamma^n_\mathsf{A})$ for which $M \vDash_{\mathbf{L}_\mathcal{E}} p_i$ iff $a_i \in E$,

   b) and vice versa, for each model $M \in \mathcal{M}_{\mathbf{L}_\mathcal{E}}(\Gamma^n_\mathsf{A})$ there is an $\mathcal{E}$-extension E of A for which $M \vDash_{\mathbf{L}_\mathcal{E}} p_i$ iff $a_i \in E$.

   Thus, the models of the logic correspond exactly to the $\mathcal{E}$-extensions in the sense that they validate exactly the propositional letters representing the arguments in the extensions.

For instance, an adequate logic for preferred extensions for our Example 7.2.1 would have two types of models: one validating p_1 and p_4, representing the preferred extension $\{a_1, a_4\}$, and another one validating p_2 and p_4, representing the preferred extension $\{a_2, a_4\}$. For skeptical acceptance, the only propositional letter derivable is supposed to be p_4, since the only skeptically acceptable argument is a_4. Due to completeness, $p_1 \vee p_2$ is obviously also a consequence. This is also intuitive since in every preferred extension either a_1 or a_2 is valid.

Definition 7.3.2. Let $\mathcal{E}$ be an extension type. If a logic $\mathbf{L}$ fulfills requirement (1) for $\mathcal{E}$ and for all AFs $\mathsf{A} = \langle \mathcal{A}, \rightarrow \rangle$ where $\mathcal{A} \subseteq \mathcal{A}_n$, then we say that $\mathbf{L}$ *syntactically represents extension type $\mathcal{E}$ with respect to skeptical (resp. credulous) acceptance for argumentation frameworks with at most n arguments.*
If a logic $\mathbf{L}$ fulfills the requirements in (2) for $\mathcal{E}$ and for all AFs $\mathsf{A} = \langle \mathcal{A}, \rightarrow \rangle$ where $\mathcal{A} \subseteq \mathcal{A}_n$, then we say that $\mathbf{L}$ *semantically represents extension type $\mathcal{E}$ for argumentation frameworks with at most n arguments.*

7.3.4 ALs – Interpreting a Premise Set as Normally as Possible

The logics we are going to present belong to the class of adaptive logics (ALs). The essential feature of ALs is that they interpret a premise set "as normally

as possible" given certain criteria for normality. Adaptive logics are defined on the basis of a so-called lower limit logic (**LLL**). The **LLL** is a a reflexive, transitive, monotonic, and compact logic that has a characteristic semantics and contains classical logic. Semantically speaking, ALs select from all **LLL**-models the ones that "are normal enough" and hence satisfy a certain standard of normality. We will give a technically precise explication of this in a moment. Translated in the context of our application and given an AF $\mathsf{A} = \langle \mathcal{A}, \rightarrow \rangle$ this means, for instance,

- that, in the case of preferred extensions, in each selected **LLL**-model of Γ^n_{A} as many arguments as possible[7] are validated while the criteria for admissible extensions are satisfied,
- that, in the case of grounded extensions, in each selected **LLL**-model of Γ^n_{A} as few arguments as possible are validated while the criteria for complete extensions are satisfied,
- that, in the case of semi-stable extensions, in each selected **LLL**-model of Γ^n_{A} as many arguments as possible are validated and at the same time as many arguments as possible are defeated while the criteria for complete extensions are satisfied.

An adaptive **strategy** defines what it means for a model to satisfy the standard of normality. For all the ALs in this chapter it will mean that the models should be as "normal as possible", or in other words, "minimally abnormal". So-called **abnormalities** define what is considered as abnormal for these models. Abnormalities are a certain set of formulas defined by a logical form. For instance, in case of preferred extension

$$\Omega_P = \{\neg p_i \mid i \leq n\}$$

is a good choice for abnormalities. Assume for the moment that, given an AF A, the **LLL**-models of Γ^n_{A} correspond to the admissible extensions of A.[8] The minimal abnormal models are the **LLL**-models in which as few negated propositional letters $\neg p_i$ are validated as possible. Inversely that means that as many propositional letters are validated as possible. Then these models correspond exactly to the preferred extensions, since these are the maximal admissible extensions.

More generally, where Ω is the set of abnormalities and $\mathbf{L}$ is a logic, we define the abnormal part of an $\mathbf{L}$-model M as follows, $\mathrm{Ab}^{\mathbf{L}}_{\Omega}(M) = \{\varphi \in \Omega \mid M \vDash_{\mathbf{L}} \varphi\}$. We say that an $\mathbf{L}$-model M of Γ is an Ω-*minimally abnormal* $\mathbf{L}$-*model of* Γ iff for all $\mathbf{L}$-models M' of Γ, $\mathrm{Ab}^{\mathbf{L}}_{\Omega}(M') \not\subset \mathrm{Ab}^{\mathbf{L}}_{\Omega}(M)$. The ALs presented in this chapter

[7] More precisely we would have to express this by "as many propositional letters as possible that represent arguments of the given AF".

[8] Actually, as we will see in Section 7.3.5, the $\mathbf{L_A}$-models of Γ^n_{A} are a superset of the models corresponding to the admissible extensions of A. Due to this we will perform a preselection on the $\mathbf{L_A}$-models before selecting the Ω_P-minimally abnormal models (see Section 7.3.6).

select all the Ω-minimally abnormal **LLL**-models of a given premise set where the exact nature of Ω depends on the extension type under consideration.

In terms of proofs, this idea is realized by allowing for certain lines to be added to the proof conditionally while the **LLL**-rules are unconditionally applicable. For instance in case of preferred extensions we are interested in adding an argument α on the condition that the assumption that $\neg\alpha$ is not the case is safe. The adaptive strategy informs the logic when such a condition should be considered as unsafe. In that case the line with the unsafe condition is marked as invalid.

But let us exemplify the notions just introduced by having a look at a proof for the simple AF given by $\mathsf{A}^1 = \langle\{a_1, a_2\}, \rightarrow^1\rangle$ where $\rightarrow^1 = \{(a_1, a_2)\}$. As discussed above, the premise set corresponding to A^1 for language $\mathcal{W}_3$ is given by $\Gamma^3_{\mathsf{A}^1} = \{p_1 \twoheadrightarrow p_2, \bot \twoheadrightarrow p_1, \bot \twoheadrightarrow p_2\}$. We use more than two propositional letters (namely three) since we are later going to enhance A^1. The only preferred extension for A^1 is $\{a_1\}$. Thus, what is expected from the logic for preferred extensions is to derive p_1. A good choice for an **LLL** is our core system $\mathbf{L_A}$.

1	$p_1 \twoheadrightarrow p_2$	PREM	$\varnothing$
2	p_1	RC	$\{\neg p_1\}$
3	$\neg p_2$	1,2; R$\twoheadrightarrow$	$\{\neg p_1\}$

Note that the adaptive proof is enhanced by a forth column stating the condition of a line. Clearly, we don't need any conditions in order to introduce premises, as it is done at line 1. We write PREM for the generic rule allowing to introduce premises on the empty condition.

As discussed above, the idea for preferred extensions is to derive as many arguments as possible. Technically, this is made possible by allowing for the conditional introduction of arguments, i.e., propositional letters. For instance at line 2, p_1 is introduced on the condition $\{\neg p_1\}$. The elements of conditions are abnormalities, in our case members of Ω_P.

Now, once p_1 is considered to be valid, we know that p_2 cannot be valid since $p_1 \twoheadrightarrow p_2$. Indeed, at line 3 we derive $\neg p_2$ by rule (R$\twoheadrightarrow$). Note that the condition of line 2 is carried forward to line 3 since the derivational step performed at line 3 employs also line 2.

The following two generic rules for ALs characterize what has just been demonstrated in our example:

RU If $\varphi_1, \ldots, \varphi_n \vdash_{\mathbf{LLL}} \psi$:

$$\begin{array}{ll} \varphi_1 & \Delta_1 \\ \vdots & \vdots \\ \varphi_n & \Delta_n \\ \hline \psi & \Delta_1 \cup \cdots \cup \Delta_n \end{array}$$

$$\text{RC}\quad \text{If } \varphi_1,\ldots,\varphi_n \vdash_{\mathbf{LLL}} \psi \vee \mathsf{Dab}(\Theta):\quad \begin{array}{ll} \varphi_1 & \Delta_1 \\ \vdots & \vdots \\ \varphi_n & \Delta_n \\ \hline \psi & \Delta_1 \cup \cdots \cup \Delta_n \cup \Theta \end{array}$$

In the case of RC, Θ is a finite and non-empty set of abnormalities. We write $\mathsf{Dab}(\Theta)$ to abbreviate $\bigvee_{\varphi\in\Theta}\varphi$. RU states that if a formula ψ is derived on a line l by an **LLL**-rule with antecedents that are derived on conditions $\Delta_1,\ldots,\Delta_n$, then these conditions are carried forward to line l. In our example we derive $\neg p_2$ at line 3 by the **LLL**-rule (R$\twoheadrightarrow$) with antecedents in lines 1 and 2. The conditions of these lines, namely $\varnothing$ and $\{\neg p_1\}$, are carried forward to line 3.

The essential strength of adaptive logics comes with rule RC. It enables us to derive formulas conditionally. Since $\vdash p_1 \vee \neg p_1$ is a theorem of propositional logic, we derive p_1 by RC at line 2 on the condition $\{\neg p_1\}$. Also for applications of RC conditions are carried forward, as it was the case for RU.

Note that adaptive proofs are not yet fully characterized by the generic rules PREM, RU and RC. What is missing are means to invalidate lines which are derived on conditions that have to be considered as unsafe. The marking definition of our adaptive logics will give a precise account of when a condition is considered as unsafe. We will come to that at the end of the next subsection.

7.3.5 The Problem of an Interpretative Surplus

It was indicated above that an intuitive semantic selection procedure is to select all Ω_P-minimally abnormal $\mathbf{L_A}$-models of $\Gamma^3_{\mathsf{A}^1}$. This section will show that this idea, although it is on the right path, gives rise to a problem. Namely, some of the models selected by the proposed procedure validate attacks, $p_i \twoheadrightarrow p_j$, and propositional letters, p_k, that do not correspond to attacks or arguments in the given AF A, i.e., $(a_i, a_j) \notin \rightarrow$ and $a_k \notin \mathcal{A}$. Thus, some models 'interpret too much into the given AF'. After explicating the problem in this subsection we will propose a solution by refining our semantic selection (Subsection 7.3.6).

Note that, in our example, neither is a_1 a member of all admissible extensions of A nor is p_1 derivable by $\mathbf{L_A}$. Argument a_1 constitutes the unique preferred extension $\{a_1\}$ and thus p_1 should be derivable by a logic for preferred extensions and furthermore, it should be the only propositional letter derivable. Also, since $(p_1 \twoheadrightarrow p_2) \supset (\neg p_1 \vee \neg p_2)$, p_1 and p_2 are never both valid in the same model. This is as desired since a_1 attacks a_2. Furthermore, it can be easily shown that there is an Ω_P-minimally abnormal $\mathbf{L_A}$-model of Γ^3_{A} that verifies p_1 and $\neg p_2$. There are however two problems:

(1) As can easily be verified, all the Ω_P-minimally abnormal $\mathbf{L_A}$-models of $\Gamma^3_{\mathsf{A}^1}$ also validate p_3. Since a_3 is not part of the AF in question, A^1, this is undesired and is not in accordance with our adequacy requirements.

(2) It is easy to see that there are other Ω_P-minimally abnormal $\mathbf{L_A}$-models which verify p_2 and $\neg p_1$. Some of these models verify $p_2 \nrightarrow p_1$, which enables p_2 to defend itself against the attack from p_1. However, a_2 does not attack a_1 in our AF A^1.

Thus, the problem is that, in order to validate as many arguments as possible, the logic selecting Ω_P-minimally abnormal $\mathbf{L_A}$-models, (a), validates propositional letters that do not correspond to arguments of the given AF, and, (b), in some models validates attacks which are not part of the given premises. To interpret a premise set as normally as possible the logic should thus always also take care, (a), that all the arguments that do not correspond to arguments in the given AF are not validated, and, (b), that no additional attacks are derived or validated in the models.

One way to do so would be to directly enhance the premise set for AF $\langle \mathcal{A}, \rightarrow \rangle$ by $p_i \nrightarrow p_j$ iff $(a_i, a_j) \notin \rightarrow$ and by $\bot \nrightarrow p_i$ iff $a_i \notin \mathcal{A}$. In our case the enriched premise set is

$$\Gamma'_{\mathsf{A}^1} = \Gamma^3_{\mathsf{A}^1} \cup \bigcup_{i=1}^{3}\{p_i \nrightarrow p_1\} \cup \bigcup_{i=2}^{3}\{p_i \nrightarrow p_2\} \cup \bigcup_{i=1}^{3}\{p_i \nrightarrow p_3\} \cup \{\bot \nrightarrow p_3\}.$$

However, to enhance the premise set in this way has disadvantages. Some of them are rather obvious: our first proposal for the representation of AFs in terms of premise sets as exemplified by $\Gamma^3_{\mathsf{A}^1}$ is more intuitive, simple and elegant compared to the enhanced presentation exemplified by Γ'_{A^1}. Furthermore, such enhanced premise sets can have a very high cardinality (namely, $n! + n$) already for small AFs. Instead of manually adding all the additional formulas to the premise set $\Gamma^3_{\mathsf{A}^1}$ it would be better if the logic were able to derive them on its own.

Additionally, for some applications the enhancement of the premise set proposed above is counterproductive. Suppose we are to model the argumentative reasoning of intelligent agents: argumentation is a dynamic process, new information in form of new arguments and new attack relations might come in (see also (Cayrol et al. 2008)). In terms of argumentation frameworks that means that the initial state of an argumentation might be given by A^1, while, at a later stage a new argument might enter the scene. For instance an agent might, in order to defend a_2 argue that a new argument a_3 attacks a_1. Thus, A^1 is extended to $\mathsf{A}^2 = \langle \{a_1, a_2, a_3\}, \{(a_1, a_2), (a_3, a_1)\} \rangle$. Now if we represent A^1 by the premise set Γ'_{A^1}, then there is no way anymore to introduce $p_3 \nrightarrow p_1$ at a later point of the proof, since this contradicts the premise $p_3 \nrightarrow p_1 \in \Gamma'_{\mathsf{A}^1}$. Were we to add $p_3 \nrightarrow p_1$, this would lead to explosion. Similar applications requiring from the logic the ability to deal with new information on-the-fly would be in the fields of belief revision or machine learning (see page 125). Fortunately, ALs offer a way to avoid these difficulties. The following subsection explores how.

7.3.6 A Better Solution: Going Adaptive and Enabling External Dynamics

Instead of enriching the premise set in the way demonstrated above, it would thus be preferable to have a logic that, (a), has the virtue of dealing with such cases of external dynamics, that is to say a logic which is able to deal with the addition of new arguments and new attacks at any point during the proof without exhibiting explosive behaviour, and, (b), can deal with the intuitive and simple representation Γ^n_{A} of AFs A as premise sets as defined in Section 7.3.2 without exhibiting the problems explicated in Subsection 7.3.5.

This is where the strengths of adaptive logics can again be of use. The idea is now to interpret the relation between two arguments a_i and a_j as non-attacking as long as the premise $p_i \twoheadrightarrow p_j$ has not been introduced, and to treat p_i as not representing an argument as long as the premise $\bot \twoheadrightarrow p_i$ has not been introduced. For our example that means that as long as our agent doesn't introduce p_3, the logic should, (a), treat the relation between p_3 and p_i for all $p_i \in \mathcal{V}_n$ as non-attacking and thus derive $p_3 \not\twoheadrightarrow p_i$, and, (b), derive $\bot \not\twoheadrightarrow p_3$ and hence by (R$\bot$) $\neg p_3$. As a result, as long as p_3 has not been introduced, the only argument in the consequence set should be p_1 since a_1 constitutes the unique preferred extension of A^1. However, as soon as p_3 and $p_3 \twoheadrightarrow p_1$ have been introduced, we are interested in deriving p_2 and p_3 as only arguments. The reason is that $\{a_2, a_3\}$ constitutes the unique preferred extension of A^2. As the reader might have already guessed, the way to achieve this behaviour via an adaptive logic is to define abnormalities by the logical form $\alpha \twoheadrightarrow \beta$. Let thus

$$\Omega_{\twoheadrightarrow} =_{\text{df}} \{\alpha \twoheadrightarrow \beta \mid \alpha \in \mathcal{V}_n \cup \{\bot\}, \beta \in \mathcal{V}_n\}$$

The idea is to ensure in this way that $p_i \not\twoheadrightarrow p_j$ and $\bot \not\twoheadrightarrow p_k$ are derivable whenever $p_i \twoheadrightarrow p_j$ and $\bot \twoheadrightarrow p_k$ are not part of the premise set.

We have seen in Subsection 7.3.5 that the Ω_P-minimally abnormal $\mathbf{L_A}$-models of $\Gamma^3_{\mathsf{A}^1}$ do not correspond to the preferred extensions since they validate arguments and attacks which are not part of A^1. In order to improve on that we employ a so-called *prioritized adaptive logic*. In semantic terms the idea is realized in two steps:

(1) First, we pre-select the set of $\mathbf{L_A}$-models $\mathcal{M}_{\twoheadrightarrow}$ of $\Gamma^3_{\mathsf{A}^1}$ which validate only the attacks that are actually a part of the given AF and which invalidate all propositional letters that do not represent arguments.

(2) Second, from our preselection $\mathcal{M}_{\twoheadrightarrow}$ we select the Ω_P-minimally abnormal $\mathbf{L_A}$-models.

Now we have all tools at hand, at least semantically, to introduce our logic for preferred extensions:

$$\mathbf{AL_P} = \langle \mathbf{L_A}, [\Omega_{\twoheadrightarrow}, \Omega_P], [\text{simple strategy}, \text{minimal abnormality strategy}] \rangle$$

The first element, $\mathbf{L_A}$, is the lower limit logic. The second element lists the abnormalities for the first and second selection. The third element lists the strategies used for the semantic selections, or syntactically, for the markings in the proof. We will comment on the simple strategy more in a moment; what is now important is that semantically both strategies select minimally abnormal models.[9] $\mathbf{AL_P}$ is semantically characterized by the two steps of the selection procedure which have just been introduced. Only performing the first step characterizes another, flat adaptive logic that can be shown to represent admissible extensions:

$$\mathbf{AL_A} = \langle \mathbf{L_A}, \Omega_{\twoheadrightarrow}, \text{simple strategy} \rangle$$

It is easy to prove that, for a given AF $\mathsf{A} = \langle \mathcal{A}, \rightarrow \rangle$, the models selected by the first selection, i.e., the $\mathbf{AL_A}$-models of Γ^n_{A},

(a) validate $p_i \twoheadrightarrow p_j$ iff $(a_i, a_j) \in \rightarrow$, and,

(b) for all $a_i \notin \mathcal{A}$ validate $\bot \nrightarrow p_i$ and thus, by (s$\bot$), $\neg p_i$.

This obviously solves the problem of Subsection 7.3.5. Indeed, the $\mathbf{AL_A}$-models of $\Gamma^3_{\mathsf{A}^1}$ correspond to the admissible extensions of A^1.[10] This is as expected, since the models in our second selection, the Ω_P-minimally abnormal models from these $\mathbf{AL_A}$-models, are expected to correspond to the maximal admissible sets. We will see in Section 7.4 that the same sequential selection procedure is applied to other extension types, only the abnormalities for the second selection have to be adjusted.

We have talked a lot about semantics. Let us now take a look at a continuation of our proof from page 128 and see how the ideas presented above are applied syntactically.

$^{13}4$	p_2	RC	$\{\neg p_2\}$

[9] The reader should not be confused by the fact that for both strategies, simple strategy resp. minimal abnormality, we apply the same semantic selection, namely the selection of minimally abnormal $\mathbf{L_A}$-models with respect to the abnormalities in $\Omega_{\twoheadrightarrow}$ resp. Ω_P. The reason for this is that the simple strategy is equivalent to the minimal abnormality strategy for a lower limit logic **LLL**, abnormalities Ω and a class of premise sets $\mathbf{\Gamma}$ if the following fact holds:

> ($F\star$): For all $\Gamma \in \mathbf{\Gamma}$ and all finite and non-empty $\Delta \subseteq \Omega$, $\Gamma \vdash_{\mathbf{LLL}} \mathsf{Dab}(\Delta)$, then there is a $\varphi \in \Delta$ such that $\Gamma \vdash_{\mathbf{LLL}} \varphi$.

This is the case for our $\mathbf{L_A}$, $\Omega_{\twoheadrightarrow}$ and premise sets defined by Γ^n_{A} (as shown in (Straßer and Šešelja 2009)). Hence, in this case the simple strategy, as we will see, allows for a simplified marking strategy (see Definition 7.3.3) compared to the one for minimal abnormality (which is defined in Subsection 7.3.8, Definition 7.3.4). Of course, due to ($F\star$) the semantic selection for the simple strategy can also be characterized as follows: selected are all $\mathbf{L_A}$-models of Γ^n_{A} that validate only those abnormalities in $\Omega_{\twoheadrightarrow}$ that are $\mathbf{L_A}$-derivable from Γ^n_{A} (or equivalently, that are validated by all other $\mathbf{L_A}$-models of Γ^n_{A}). Note, that in the case that fact ($F\star$) does not hold, such models are not guaranteed to exist. See also the discussion in (Straßer 2010, Section 2.4.3).

[10] The representational results are stated in Section 7.4 (see Theorem 7.4.1 and Corollary 7.4.1) and proven in (Straßer and Šešelja 2009).

$^{13}5$	$\mathsf{def}\, p_1$	1,4; Rad	$\{\neg p_2\}$
$^{13}6$	$\neg p_1$	1,4; R$\nsubseteq$	$\{\neg p_2\}$
$^{13}7$	$\bigvee_{i=1}^{3}(p_i \wedge (p_i \nsubseteq p_1))$	5; Def	$\{\neg p_2\}$
$^{13}8$	$\left[\begin{array}{l}(p_2 \wedge (p_2 \nsubseteq p_1))\vee \\ \quad (p_3 \wedge (p_3 \nsubseteq p_1))\end{array}\right.$	6,7; RU	$\{\neg p_2\}$
$^{15}9$	p_3	RC	$\{\neg p_3\}$
10	$\left[\begin{array}{l}\neg p_2 \vee (p_2 \wedge (p_2 \nsubseteq p_1)) \\ \quad \vee (p_3 \wedge (p_3 \nsubseteq p_1))\end{array}\right.$	8; RU	$\varnothing$
11	$p_2 \not\nsubseteq p_1$	RC	$\{p_2 \nsubseteq p_1\}$
12	$p_3 \not\nsubseteq p_1$	RC	$\{p_3 \nsubseteq p_1\}$
13	$\neg p_2$	10,11,12; RU	$\{p_2 \nsubseteq p_1, p_3 \nsubseteq p_1\}$
14	$\bot \not\nsubseteq p_3$	RC	$\{\bot \nsubseteq p_3\}$
15	$\neg p_3$	14; R$\bot$	$\{\bot \nsubseteq p_3\}$

What is happening in the proof segment above? At line 4 we conditionally introduce p_2. This gives rise to p_1 being defeated under the same condition at line 5. Furthermore, at line 8 we derive that either p_2 or p_3 has to be the defeater of p_1. Moreover, we introduce p_3 conditionally at line 9. What we expect from the proof is that lines 4–9 get invalidated, since it is not in our interest to derive p_2 and p_3, as neither is a part of the unique preferred extension $\{a_1\}$.

At lines 11, 12 and 14 we realize the ideas from above, namely that two propositional letters p_i and p_j should be considered to not attack each other as long as no premise $p_i \nsubseteq p_j$ has been introduced, and that a propositional letter p_i should be considered as not valid unless one of the introduced premises states that it is part of the AF under consideration, i.e., $\bot \nsubseteq p_i$. Thus, we derive $p_2 \not\nsubseteq p_1$, $p_3 \not\nsubseteq p_1$ and $\bot \not\nsubseteq p_3$ conditionally at line 11, 12 and 14. Since $p_2 \nsubseteq p_1$, $p_3 \nsubseteq p_1$ and $\bot \nsubseteq p_3$ are not part of the premise set $\Gamma^3_{\mathsf{A}^1}$, these lines are not going to be marked in our proof for A^1. Using these lines we are able to derive $\neg p_2$ at line 13 on the condition $\{p_2 \nsubseteq p_1, p_3 \nsubseteq p_1\}$ as well as $\neg p_3$ at line 15 on condition $\{\bot \nsubseteq p_3\}$. Now something very important happens. Note that lines 4–8 have been derived on the very condition that $\neg p_2$ is not valid. However, now we have derived $\neg p_2$ and thus all lines which were derived on this condition should be considered as invalid derivations and thus have to be marked. Similarly, by introducing $\bot \not\nsubseteq p_3$ at line 14 we derive $\neg p_3$ at line 15 which causes the marking of line 9. The idea behind the marking is thus to invalidate lines on conditions that have to be considered as unsafe. What is considered as unsafe depends on the adaptive strategy used. Recall that we have defined two types of abnormalities: Ω_P and $\Omega_{\nsubseteq}$. Each of these come with their own marking definition. The marking definition for abnormalities in $\Omega_{\nsubseteq}$ is very simple, after all it is based on the adaptive strategy called the *simple strategy*.

Definition 7.3.3 (Marking for the simple strategy)**.** A line with condition Δ is marked at stage s if a $\alpha \nsubseteq \beta \in \Delta \cap \Omega_{\nsubseteq}$ has been derived on the empty

condition.

Suppose for a moment that new information comes in: one agent, in order to defend a_2, voices a_3 which attacks a_1. In this case we would introduce $p_3 \twoheadrightarrow p_1$ and $\bot \twoheadrightarrow p_3$ by PREM. Note that lines 12–15 would get marked. These lines would not anymore be considered to be derived since they rely on the condition that $p_3 \twoheadrightarrow p_1$ and resp. $\bot \twoheadrightarrow p_3$ are not derivable. This behavior is obviously intuitive.

The marking conditions for Ω_P are technically a bit more complicated. We will introduce them later in Subsection 7.3.8 in order not to complicate things more than necessary at this point. However, let us make another important remark.

So far we have discussed the prioritized aspect of adaptive logics only in terms of the semantic selection. Of course, this has a syntactic equivalent to it. This is illustrated in the proof, for instance, at line 13: here we derive an abnormality $\neg p_2 \in \Omega_P$ at an unmarked line on a condition $\{p_2 \twoheadrightarrow p_1, p_3 \twoheadrightarrow p_1\} \subset \Omega_{\twoheadrightarrow}$. This causes the marking of all lines that have $\neg p_2$ as a part of the condition. Similarly, at line 15 we derive $\neg p_3$ on the condition $\{\bot \twoheadrightarrow p_3\}$ which causes the marking of line 9. Hence, lines are considered as (un)safe due to conditions in Ω_P on the basis of abnormalities in Ω_P (and their disjunctions, as we will see in Subsection 7.3.8) derived on unmarked lines on the empty condition or on conditions which are subsets of $\Omega_{\twoheadrightarrow}$. In contrast, the marking condition for $\Omega_{\twoheadrightarrow}$ requires that, in order to mark a line with condition Δ, a $\alpha \twoheadrightarrow \beta \in \Delta$ has to be derived on the empty(!) condition. It is not enough to derive $\alpha \twoheadrightarrow \beta$ on a condition $\Delta' \subseteq \Omega_P$. This evidently mirrors syntactically the prioritized aspect of the two semantic selections.

7.3.7 External Dynamics — Letting New Information In

Let us now proceed from A^1 to A^2. The new information in $\Gamma^3_{\mathsf{A}^2} \setminus \Gamma^3_{\mathsf{A}^1}$ is introduced at lines 16 and 17. We restate lines 2, 4, 9, 13–15. Let $\Theta = \{p_1 \twoheadrightarrow p_3, p_2 \twoheadrightarrow p_3, p_3 \twoheadrightarrow p_3\}$.

$^{22}2$	p_1	RC	$\{\neg p_1\}$
⋮	⋮	⋮	⋮
4	p_2	RC	$\{\neg p_2\}$
⋮	⋮	⋮	⋮
9	p_3	RC	$\{\neg p_3\}$
⋮	⋮	⋮	⋮
$^{16}13$	$\neg p_2$	10,11,12; RU	$\{p_2 \twoheadrightarrow p_1, p_3 \twoheadrightarrow p_1\}$
$^{17}14$	$\bot \nrightarrow p_3$	RC	$\{\bot \twoheadrightarrow p_3\}$
$^{17}15$	$\neg p_3$	14; R$\bot$	$\{\bot \twoheadrightarrow p_3\}$
16	$p_3 \twoheadrightarrow p_1$	PREM	$\varnothing$
17	$\bot \twoheadrightarrow p_3$	PREM	$\varnothing$
18	$\mathsf{def}\, p_3$	2,16; Rad	$\{\neg p_1\}$
19	$\bigvee_{i=1}^{3}(p_i \wedge (p_i \twoheadrightarrow p_3))$	18; Def	$\{\neg p_1\}$

20	$\neg p_1 \vee \bigvee_{i=1}^{3}(p_i \wedge (p_i \nsim\!\!\!\!\!\!\rightarrow p_3))$	19; RA	$\varnothing$
21	$\bigwedge_{i=1}^{3}(p_i \nsim\!\!\!\!\!\!\rightarrow p_3)$	RC	Θ
22	$\neg p_1$	20,21; RU	Θ

Due to the new information in A^2, p_3 now corresponds to an argument, namely a_3. Furthermore a_3 attacks a_1. Thus we introduce at lines 16 and 17 the new premises $p_3 \nsim\!\!\!\!\!\!\rightarrow p_1$ and $\bot \nsim\!\!\!\!\!\!\rightarrow p_3$. This immediately leads to the marking of line 13 since this line was derived on the condition that p_3 does not attack p_1, and to the marking of lines 14 and 15 since these lines were derived on the condition $\{\bot \nsim\!\!\!\!\!\!\rightarrow p_3\}$, i.e., that a_3 does not belong to the AF in question. Furthermore, the new information enables us to derive $\neg p_1$ at line 22 on the condition $\Theta \subset \Omega_{\nsim\!\!\!\!\!\!\rightarrow}$, which leads to the marking of line 2. Moreover, line 4 is now unmarked, since the line which caused it to be marked before, namely 13, is now itself marked. Analogously for line 9: line 15 which caused it to be marked is now marked itself and hence, 9 is unmarked.

7.3.8 The Minimal Abnormality Strategy and Final Derivability

We postponed the exact marking definition for the minimal abnormality strategy so far. Recall that this strategy is used for the abnormalities in Ω_P. The marking definition can be better motivated if we take a look at another simple example. Let $\mathsf{A} = \langle\{a_1, a_2\}, \{(a_1, a_2), (a_2, a_1)\}\rangle$. The two preferred extensions are $\{a_1\}$ and $\{a_2\}$. Thus, there should be two types of Ω_P-minimally abnormal $\mathbf{AL_A}$-models: on the one hand a model verifying p_1 and $\neg p_2$ and on the other hand a model verifying p_2 and $\neg p_1$. That means that we expect from our logic to derive $p_1 \vee p_2$ since either p_1 or p_2 is valid in each minimally abnormal model. Let us take a look at a proof for $\Gamma^4_{\mathsf{A}} = \{p_1 \nsim\!\!\!\!\!\!\rightarrow p_2, p_2 \nsim\!\!\!\!\!\!\rightarrow p_1, \bot \nsim\!\!\!\!\!\!\rightarrow p_1, \bot \nsim\!\!\!\!\!\!\rightarrow p_2\}$ and the language $\mathcal{W}_4$.

1	$p_1 \nsim\!\!\!\!\!\!\rightarrow p_2$	PREM	$\varnothing$
2	$p_2 \nsim\!\!\!\!\!\!\rightarrow p_1$	PREM	$\varnothing$
3	$\bot \nsim\!\!\!\!\!\!\rightarrow p_1$	PREM	$\varnothing$
4	$\bot \nsim\!\!\!\!\!\!\rightarrow p_2$	PREM	$\varnothing$
5	p_1	RC	$\{\neg p_1\}$
6	p_2	RC	$\{\neg p_2\}$
7	$p_1 \vee p_2$	5; RU	$\{\neg p_1\}$
8	$p_1 \vee p_2$	6; RU	$\{\neg p_2\}$

With the analyses given above, what we expect from our logic is that lines 5 and 6 are marked, while line 7 and 8 are not marked and hence $p_1 \vee p_2$ is considered as being derived. Note that with line 1 we can derive the following disjunction of abnormalities:

9	$\neg p_1 \vee \neg p_2$	1; R$\nsim\!\!\!\!\!\!\rightarrow$	$\varnothing$

It is important to notice that neither $\neg p_1$ nor $\neg p_2$ can be derived on a condition $\Delta \subseteq \Omega_{\twoheadrightarrow}$ (including the empty condition $\varnothing$).

In order to define our marking conditions we need to introduce some terminology. Let us do this in a more general setting for a logic

$$\mathbf{AL} = \langle \mathbf{L_X}, [\Omega_{\twoheadrightarrow}, \Omega], [\text{simple strategy}, \text{minimal abnormality strategy}] \rangle,$$

where $\mathbf{X} \in \{\mathbf{A}, \mathbf{C}\}$. Obviously our $\mathbf{AL_P}$ is such a logic for $\Omega = \Omega_P$. Where $\Delta \subseteq \Omega$ and Δ is finite and non-empty, we say that $\mathsf{Dab}(\Delta)$ is a Ω-*minimal Dab-formula at a stage* s of the proof iff it is the formula of an unmarked line with a condition $\Delta_{\twoheadrightarrow} \subseteq \Omega_{\twoheadrightarrow}$ and no $\mathsf{Dab}(\Delta')$, where $\Delta' \subset \Delta$, is the formula of an unmarked line with a condition $\Delta'_{\twoheadrightarrow} \subseteq \Omega_{\twoheadrightarrow}$. A *choice set* of $\Sigma = \{\Delta_1, \Delta_2, \dots\}$ is a set that contains an element out of each member of Σ. A *minimal choice set* of Σ is a choice set of Σ of which no proper subset is a choice set of Σ.[11] Where $\mathsf{Dab}(\Delta_1), \dots, \mathsf{Dab}(\Delta_n)$ are the Ω-minimal Dab-formulas at stage s for a premise set Γ, $\Phi_s(\Gamma)$ is the set of minimal choice sets of $\{\Delta_1, \dots, \Delta_n\}$.

With this terminology we can define marking conditions for the abnormalities Ω and the minimal abnormality strategy. Let Γ be a premise set.

Definition 7.3.4 (Marking for the minimal abnormality strategy (with respect to Ω))**.** Line i is marked at stage s if, where φ is derived on the condition Δ at line i,

(i) there is no $\Delta' \in \Phi_s(\Gamma)$ such that $\Delta' \cap \Delta = \varnothing$, or

(ii) for some $\Delta' \in \Phi_s(\Gamma)$, there is no line at which φ is derived on a condition Θ for which $\Delta' \cap \Theta = \varnothing$.

Let us return to our example. Note that at this stage of the proof $\neg p_1 \vee \neg p_2$ at line 9 is a Ω_P-minimal Dab-formula. Thus, the minimal choice sets at this stage of the proof are $\{\neg p_1\}$ and $\{\neg p_2\}$. By (ii) lines 5 and 6 are marked. This is as desired, since after all, neither is a_1 nor is a_2 a skeptically accepted argument. However, the situation is different for lines 7 and 8. Note that neither (i) nor (ii) apply, due to the fact that we are able to derive $p_1 \vee p_2$ on condition $\{\neg p_1\}$ and(!) on condition $\{\neg p_2\}$.

Since, as we have seen with our examples, markings come and go in adaptive proofs, we need a stable criterion for derivability in order to define a consequence relation.

Definition 7.3.5. φ is *finally derived* from Γ on line i of a proof at stage s iff

(i) φ is the second element of line i,

(ii) line i is not marked at stage s and

(iii) for every extension of the proof in which line i is marked there is a further extension in which line i is unmarked.

[11] Let for instance $\Sigma = \big\{\{1,2\},\{1,3\}\big\}$. Choice sets are $\{1\},\{1,2\},\{1,3\}$, $\{2,3\}$ and $\{1,2,3\}$. Minimal are $\{1\}$ and $\{2,3\}$.

The definition states that a formula derived at an unmarked line is finally derived in the case that there is no way anymore to mark it by extending the proof. For instance in our proof above it is easy to see that $p_1 \vee p_2$ is finally derived since there is no way of extending the proof in such a way that lines 7 and 8 get marked. This is due to the fact that neither $\neg p_1$ nor $\neg p_2$ are derivable on a condition $\Delta \subseteq \Omega_{\twoheadrightarrow}$.

Let us close this section by introducing $p_1 \twoheadrightarrow p_3, p_2 \twoheadrightarrow p_3, p_3 \twoheadrightarrow p_4, \bot \twoheadrightarrow p_3, \bot \twoheadrightarrow p_4$ to our last example so that we arrive at the AF from Example 7.2.1. Hence our new premise set is $\{p_1 \twoheadrightarrow p_2, p_2 \twoheadrightarrow p_1, p_1 \twoheadrightarrow p_3, p_2 \twoheadrightarrow p_3, p_3 \twoheadrightarrow p_4, \bot \twoheadrightarrow p_1, \bot \twoheadrightarrow p_2, \bot \twoheadrightarrow p_3, \bot \twoheadrightarrow p_4\}$. In the following proof we use the abbreviations:

$$\Theta_1 = \{p_1 \twoheadrightarrow p_1, p_3 \twoheadrightarrow p_1, p_4 \twoheadrightarrow p_1\}$$
$$\Theta_2 = \{p_2 \twoheadrightarrow p_2, p_3 \twoheadrightarrow p_2, p_4 \twoheadrightarrow p_2\}$$

The proof is as follows:

10	$p_1 \twoheadrightarrow p_3$	PREM	$\varnothing$
11	$p_2 \twoheadrightarrow p_3$	PREM	$\varnothing$
12	$p_3 \twoheadrightarrow p_4$	PREM	$\varnothing$
13	$p_1 \nrightarrow p_1 \wedge p_3 \nrightarrow p_1 \wedge p_4 \nrightarrow p_1$	RC	Θ_1
14	$\mathsf{def}\, p_1 \supset p_2$	13; RU	Θ_1
15	$p_2 \nrightarrow p_2 \wedge p_3 \nrightarrow p_2 \wedge p_4 \nrightarrow p_2$	RC	Θ_2
16	$\mathsf{def}\, p_2 \supset p_1$	15; RU	Θ_2
17	$p_3 \supset \mathsf{def}\, p_1$	10; Rad	$\varnothing$
18	$p_3 \supset \mathsf{def}\, p_2$	11; Rad	$\varnothing$
19	$p_3 \supset (\mathsf{def}\, p_1 \wedge \mathsf{def}\, p_2)$	17,18; RU	$\varnothing$
20	$p_3 \supset (p_1 \wedge p_2)$	14,16,19; RU	$\Theta_1 \cup \Theta_2$
21	$\neg p_3$	9,20; RU	$\Theta_1 \cup \Theta_2$
22	p_4	RC	$\{\neg p_4\}$

It is easy to see that there is no way to mark line 22. Thus, as desired, $p_1 \vee p_2$ and p_4 are finally derivable. Recall that a_4 is the only accepted argument with respect to preferred extensions and that in each preferred extension there is either a_1 or a_2 (but obviously never both simultaneously). Thus, p_1 and p_2 are not derivable, but $p_1 \vee p_2$ and $\neg p_1 \vee \neg p_2$ are.

After having introduced a logic for preferred extensions it is time to introduce the other logics for abstract argumentation in the next section.

7.4 The AL framework for Skeptical Acceptance

In this section we will introduce adaptive logics for all the standard extension types for abstract argumentation. After our discussion in the previous section this can be done very smoothly. Let us first recapitulate the three characteristic elements of ALs that were introduced in the previous section:

The lower limit logic (LLL) While all the rules of **LLL** are valid in **AL**, the latter additionally allows for certain rules to be applied conditionally. This strengthens **LLL** as it allows to derive at least as much as **LLL** and in most cases even more. Thus, the consequence set of **AL** is a superset of the consequence set of **LLL**: $Cn_{\mathbf{AL}}(\Gamma) \supseteq Cn_{\mathbf{LLL}}(\Gamma)$ for all premise sets Γ. In semantic terms ALs select a subset of "sufficiently normal" **LLL**-models. In the case of our logics, models are "minimally abnormal" in the sense that they validate as few abnormalities as possible. This brings us to the next point:

The abnormalities The set of abnormalities is defined by a logical form F.[12] In the last section we, for instance, used abnormalities of the form $\neg\alpha$ where α is a propositional letter. Thus, one set of abnormalities was characterized by the set $\Omega_P = \{\neg p_i \mid i \leq n\}$ for the language $\mathcal{W}_n$.

The strategy Together with the abnormalities the strategy gives an exact account of what it means to interpret a premise set "as normally as possible". In the previous section we, for instance, employed the minimal abnormality strategy. In semantic terms this strategy selects all **LLL**-models M of a given premise set Γ for which there are no **LLL**-models that validate less abnormalities (w.r.t. $\subset$). In syntactic terms strategies are realized by a marking definition. In the last section we have demonstrated that adaptive proofs are dynamic: the markings invalidate lines, however if new information is introduced and/or while we reason along, markings may come and go. In this respect adaptive proofs resemble human reasoning.

We have introduced the following notation to define adaptive logics (where $\mathbf{X} \in \{\mathbf{A}, \mathbf{C}\}$):

$$\mathbf{AL_1} = \langle \mathbf{L_X}, \Omega_{\twoheadrightarrow}, \text{simple strategy} \rangle$$
$$\mathbf{AL_2} = \langle \mathbf{L_X}, [\Omega_{\twoheadrightarrow}, \Omega], [\text{simple strategy}, \text{minimal abnormality}] \rangle$$

We have already seen examples for both cases: the logic $\mathbf{AL_A}$ is a flat AL for admissible extensions. On the other hand, the logic $\mathbf{AL_P}$ for preferred extensions is a prioritized AL.

The idea of the prioritization can be easily put in semantic terms: first the set of minimally abnormal $\mathbf{L_X}$-models $\mathcal{M}_{\twoheadrightarrow}$ with respect to $\Omega_{\twoheadrightarrow}$ is selected, and then, from these selected models in $\mathcal{M}_{\twoheadrightarrow}$ the minimally abnormal models with respect to Ω are selected. This is mirrored in the proof dynamics: the marking definition for the minimal abnormality strategy (w.r.t. Ω) is such that not only Ω-minimal Dab-formulas (i.e., minimal disjunctions of abnormalities) derived on the empty condition are considered for the marking procedure, but also such derived at unmarked lines with conditions $\Delta \subseteq \Omega_{\twoheadrightarrow}$ (see Definition 7.3.4).

We presuppose the language $\mathcal{W}_n$ throughout this and the next section for an arbitrary natural number n. Let us give a general account of the consequence

[12] F is considered to be **LLL**-contingent, i.e., neither $\vdash_{\mathbf{LLL}} \mathsf{F}$ nor $\vdash_{\mathbf{LLL}} \neg\mathsf{F}$.

relations for the flat and prioritized logics that we are going to introduce in this chapter:

Definition 7.4.1. We define $\Gamma \Vdash_{\mathbf{AL_1}} \varphi$ iff for all $\Omega_{\twoheadrightarrow}$-minimally abnormal $\mathbf{L_X}$-models M of Γ, $M \vDash_{\mathbf{L_X}} \varphi$. Furthermore, $\Gamma \vdash_{\mathbf{AL_1}} \varphi$ iff φ is finally derivable in terms of the marking conditions for the simple strategy defined in Definition 7.3.3.

We define $\Gamma \Vdash_{\mathbf{AL_2}} \varphi$ iff for all Ω-minimally abnormal $\mathbf{AL_1}$-models M of Γ, $M \vDash_{\mathbf{L_X}} \varphi$. Furthermore, $\Gamma \vdash_{\mathbf{AL_2}} \varphi$ iff φ is finally derivable in terms of the marking conditions for the simple strategy in Definition 7.3.3 and for the minimal abnormality strategy in Definition 7.3.4.

The consequence set of the prioritized logic $\mathbf{AL_2}$ in Definition 7.4.1 is characterized by $Cn_{\mathbf{AL_2}}(\Gamma) = Cn_{\mathbf{AL'_2}}(Cn_{\mathbf{AL_1}}(\Gamma))$ where[13]

$$\mathbf{AL'_2} = \langle \mathbf{LLL}, \Omega, \text{minimal abnormality strategy} \rangle$$

We are now able to define our adaptive logic framework for abstract argumentation. The idea is that we first define flat, non-prioritized adaptive logics for admissible and complete extensions. While $\mathbf{AL_A}$ is the logic for admissible extensions, the logic for complete extensions is the strengthening of $\mathbf{AL_A}$ by the rule (RCo):

$$\mathbf{AL_C} = \langle \mathbf{L_C}, \Omega_{\twoheadrightarrow}, \text{simple strategy} \rangle$$

In the second step we define the adaptive logics for all the other extension types (preferred, grounded and semi-stable) by simply adding an "adaptive layer" to the flat adaptive logics $\mathbf{AL_A}$, resp. $\mathbf{AL_C}$. We have seen this already in Section 7.3.6 for preferred extensions: we added a second level to the adaptive logic for admissible extensions resulting in logic $\mathbf{AL_P}$.

For grounded extensions the idea is similar. There are three differences compared to preferred extensions:

- While preferred extensions are a certain selection of admissible extensions, the grounded extension is a certain complete extension. Thus, instead of using $\mathbf{L_A}$ as $\mathbf{LLL}$, we now use $\mathbf{L_C}$.

- While preferred extensions were *maximal* admissible extensions, the grounded extension is the *minimal* complete extension. Thus, instead of verifying as many propositional letters as possible we now verify as few as possible. Hence, instead of defining the abnormalities as $\Omega_P = \{\neg p_i \mid i \leq n\}$ we now define them as

$$\Omega_G = \{p_i \mid i \leq n\}$$

[13] We have not characterized the marking conditions for minimal abnormality for logics that employ the minimal abnormality strategy for the flat case such as $\mathbf{AL'_2}$. They are a straightforward specification of our Definition 7.3.4.

- While there may be many preferred extensions, the grounded extension is always *unique*. This allows for a simplification, namely to use instead of the minimal abnormality strategy the simple strategy.[14]

Thus, we define the adaptive logic for grounded extensions as follows:[15]

$$\mathbf{AL_G} = \langle \mathbf{L_C}, [\Omega_{\twoheadrightarrow}, \Omega_G], [\text{simple strategy}, \text{simple strategy}] \rangle$$

The adaptive logic for semi-stable extensions shouldn't come as a surprise anymore: instead of maximizing the number of arguments validated, we now maximize not only the number of arguments validated but also the number of defeated arguments. Thus, our abnormalities are defined by

$$\Omega_S = \{\neg p_i \wedge \neg \,\mathsf{def}\, p_i \mid i \leq n\}$$

For as many propositional letters as possible the logic is supposed to derive $\neg(\neg p_i \wedge \neg \,\mathsf{def}\, p_i)$. This is equivalent to $p_i \vee \mathsf{def}\, p_i$: either p_i is valid or it is defeated. We define our adaptive logic for semi-stable extensions as follows:[16]

$$\mathbf{AL_S} = \langle \mathbf{L_C}, [\Omega_{\twoheadrightarrow}, \Omega_S], [\text{simple strategy}, \text{minimal abnormality strategy}] \rangle$$

Completeness and soundness of all the logics presented in this paper are proved in (Straßer and Šešelja 2009). The following results show that the logics defined above satisfy our representational requirements from Subsection 7.3.3.

Theorem 7.4.1. $\left\{ \begin{array}{ll} (i) & \mathbf{AL_A} \\ (ii) & \mathbf{AL_C} \\ (iii) & \mathbf{AL_P} \\ (iv) & \mathbf{AL_G} \\ (v) & \mathbf{AL_S} \end{array} \right\}$ *semantically represents* $\left\{ \begin{array}{l} \textit{admissible} \\ \textit{complete} \\ \textit{preferred} \\ \textit{grounded} \\ \textit{semi-stable} \end{array} \right\}$ *extensions for argumentation frameworks with at most n arguments.*

Corollary 7.4.1. $\left\{ \begin{array}{ll} (i) & \mathbf{AL_A} \\ (ii) & \mathbf{AL_C} \\ (iii) & \mathbf{AL_P} \\ (iv) & \mathbf{AL_G} \\ (v) & \mathbf{AL_S} \end{array} \right\}$ *syntactically represents* $\left\{ \begin{array}{l} \textit{admissible} \\ \textit{complete} \\ \textit{preferred} \\ \textit{grounded} \\ \textit{semi-stable} \end{array} \right\}$ *extensions with respect to skeptical acceptance for argumentation frameworks*

[14] As is well-known in the adaptive logic research, in case all minimally abnormal models validate the same set of abnormalities, the minimal abnormality strategy and the simple strategy are equivalent (cf. Footnote 9). See Section 2.4.3 in (Straßer 2010).

[15] In view of our discussion it is straightforward to define the marking conditions for the simple strategy for Ω_G in $\mathbf{AL_G}$: A line with condition Δ is marked at stage s if a $p_i \in \Delta \cap \Omega_G$ has been derived at an unmarked line on a condition $\Delta' \subseteq \Omega_{\twoheadrightarrow}$.

[16] In accordance with Footnote 5, $\mathbf{AL_S}$ can easily be shown to be equivalent to $\langle \mathbf{L_A}, [\Omega_{\twoheadrightarrow}, \Omega_S]$, [simple strategy, minimal abnormality strategy]$\rangle$.

with at most n arguments.

7.5 Adaptive Logics for Credulous Acceptance

So far we have presented logics modeling skeptical acceptance. The current section will deal with credulous acceptance. In the skeptical case we were interested in arguments located in the intersection of all extensions of a given type. Now we are focusing on their union. The so-called normal selections strategy will prove to be very useful for this purpose.[17] We will see that, given the systems for skeptical acceptance, everything which has to be done in order to model credulous acceptance is to use the normal selections strategy instead, resp. on top of the minimal abnormality strategy.

The reason for this can be easily understood when we take a look at the normal selections strategy from a semantic point: like the minimal abnormality strategy, the normal selections strategy selects minimally abnormal **LLL**-models. However, semantic consequences are not defined in terms of the intersection of the models but in terms of their union. This obviously mirrors the difference between skeptical and credulous acceptance, where the former is defined with respect to the intersection of all models of a certain extension type while the latter is defined in terms of the union of these extensions.

We use in this section the language $\mathcal{W}_n$ for an arbitrary natural number n. The semantic consequence relation for our prioritized logics for credulous acceptance is defined as follows:

Definition 7.5.1. Where $\mathbf{X} \in \{\mathbf{A}, \mathbf{C}\}$ let

$$\mathbf{AL^n} = \langle \mathbf{L_X}, [\Omega_{\twoheadrightarrow}, \Omega], [\text{simple strategy}, \text{normal selections}] \rangle.$$

Where $\mathcal{M}_{\mathbf{AL^n}}(\Gamma)$ is the set of all Ω-minimally abnormal $\mathbf{AL_X}$-models of Γ, we define the semantic consequence relation as follows: $\Gamma \Vdash_{\mathbf{AL^n}} \varphi$ iff there is a $M \in \mathcal{M}_{\mathbf{AL^n}}(\Gamma)$ for which $M \models_{\mathbf{L_X}} \varphi$.[18]

Thus, we are going to define, for instance, a logic for preferred extensions

[17] The normal selections strategy was first introduced in (Batens 2000). See also (Batens et al. To appear) and Section 2.9 in (Straßer 2010) for a more elaborated representation.

[18] Usually the semantic consequence relation has to be defined in terms of equivalence classes of Ω-minimally abnormal $\mathbf{AL_X}$-models. For two $\mathbf{AL_X}$-models $M \sim N$ iff $\mathrm{Ab}_\Omega^{\mathbf{L_X}}(M) = \mathrm{Ab}_\Omega^{\mathbf{L_X}}(N)$. The semantic consequence relation is then defined by $\Gamma \Vdash_{\mathbf{AL^n}} \varphi$ iff there is an Ω-minimally abnormal $\mathbf{AL_X}$-model M of Γ such that for all Ω-minimally abnormal $\mathbf{AL_X}$-models N of Γ for which $N \sim M$, $N \models_{\mathbf{L_X}} \varphi$ (see (Straßer 2010, Section 2.9)). However, the nature of our abnormalities and of our premise sets allows for the simplification in Definition 7.5.1 since it can easily be shown that for all AFs A and for all Ω-minimally abnormal $\mathbf{AL_X}$-models of Γ^n_{A}, M and N,

$$(M \sim N) \text{ iff } (\text{for all } \varphi \in \mathcal{W}_n, M \models_{\mathbf{L_X}} \varphi \text{ iff } N \models_{\mathbf{L_X}} \varphi)$$

The simplification is explicated in a more detailed way in (Straßer and Šešelja 2009).

with respect to credulous acceptance by

$$\mathbf{ALC_P} = \langle \mathbf{L_A}, [\Omega_{\twoheadrightarrow}, \Omega_P], [\text{simple strategy}, \text{normal selections}] \rangle.$$

Note that each of the selected models from Definition 7.5.1 exactly corresponds to a preferred extension.

We are still lacking a syntactic characterization of the normal selections strategy. The marking conditions are technically straightforward. The following definition covers the generic case for $\mathbf{AL^n}$ from Definition 7.5.1:

Definition 7.5.2 (Marking for normal selections (w.r.t. Ω))**.** Line i is marked at stage s if, where Δ is the condition of line i, $\mathsf{Dab}(\Delta \cap \Omega)$ has been derived at an unmarked line on a condition Δ' for which $\Delta' \cap \Omega = \varnothing$.

Definition 7.5.3. $\Gamma \vdash_{\mathbf{AL^n}} \varphi$ iff φ is finally derivable with respect to the marking conditions for simple selections (w.r.t. $\Omega_{\twoheadrightarrow}$) and for normal selections (w.r.t. Ω).

Let us again take a look at the AF A from our Example 7.2.1 for the logic $\mathbf{ALC_P}$ and with the language $\mathcal{W}_4$. Recall that the premise set is $\Gamma^4_{\mathsf{A}} = \{p_1 \twoheadrightarrow p_2, p_2 \twoheadrightarrow p_1, p_1 \twoheadrightarrow p_3, p_2 \twoheadrightarrow p_3, p_3 \twoheadrightarrow p_4, \bot \twoheadrightarrow p_1, \bot \twoheadrightarrow p_2, \bot \twoheadrightarrow p_3, \bot \twoheadrightarrow p_4\}$. In the proof we use the following abbreviation: $\Theta = \{p_1 \twoheadrightarrow p_1, p_3 \twoheadrightarrow p_1, p_4 \twoheadrightarrow p_1, p_2 \twoheadrightarrow p_2, p_3 \twoheadrightarrow p_2, p_4 \twoheadrightarrow p_2\}$.

1	$p_1 \twoheadrightarrow p_2$	PREM	$\varnothing$
2	$p_2 \twoheadrightarrow p_1$	PREM	$\varnothing$
3	$p_1 \twoheadrightarrow p_3$	PREM	$\varnothing$
4	$p_2 \twoheadrightarrow p_3$	PREM	$\varnothing$
5	$p_3 \twoheadrightarrow p_4$	PREM	$\varnothing$
6	p_1	RC	$\{\neg p_1\}$
7	$\neg p_2 \wedge \mathsf{def}\, p_2$	1,6; R$\twoheadrightarrow$,Def	$\{\neg p_1\}$
8	p_2	RC	$\{\neg p_2\}$
9	$\neg p_1 \wedge \mathsf{def}\, p_1$	2,8; R$\twoheadrightarrow$,Def	$\{\neg p_2\}$
[13]10	$p_1 \wedge \neg p_1$	6,9; RU	$\{\neg p_1, \neg p_2\}$
[13]11	$p_2 \wedge \neg p_2$	7,8; RU	$\{\neg p_1, \neg p_2\}$
[13]12	$p_1 \wedge p_2$	6,8; RU	$\{\neg p_1, \neg p_2\}$
13	$\neg p_1 \vee \neg p_2$	1; R$\twoheadrightarrow$	$\varnothing$
14	$\bigwedge_{i \in \{1,3,4\}} p_i \not\twoheadrightarrow p_1$	RC	$\{p_1 \twoheadrightarrow p_1, p_3 \twoheadrightarrow p_1, p_4 \twoheadrightarrow p_1\}$
15	$\mathsf{def}\, p_1 \supset p_2$	14; Def	$\{p_1 \twoheadrightarrow p_1, p_3 \twoheadrightarrow p_1, p_4 \twoheadrightarrow p_1\}$
16	$\bigwedge_{i \in \{2,3,4\}} p_i \not\twoheadrightarrow p_2$	RC	$\{p_2 \twoheadrightarrow p_2, p_3 \twoheadrightarrow p_2, p_4 \twoheadrightarrow p_2\}$
17	$\mathsf{def}\, p_2 \supset p_1$	16; Def	$\{p_2 \twoheadrightarrow p_2, p_3 \twoheadrightarrow p_2, p_4 \twoheadrightarrow p_2\}$
18	$p_3 \supset \mathsf{def}\, p_1$	3; Rad	$\varnothing$
19	$p_3 \supset \mathsf{def}\, p_2$	4; Rad	$\varnothing$
20	$p_3 \supset (\mathsf{def}\, p_1 \wedge \mathsf{def}\, p_2)$	18,19; RU	$\varnothing$
21	$p_3 \supset (p_1 \wedge p_2)$	15,17,20; RU	Θ
22	$\neg p_3$	13,21; RU	Θ
23	p_4	RC	$\{\neg p_4\}$

24	$\bot \twoheadrightarrow p_4$	PREM	$\varnothing$

Although the proof is very similar to the one presented in Section 7.3.8, there are important differences. Let us take a closer look. At lines 1–5 we introduce some premises. At line 6 we conditionally derive p_1. Next, due to p_1 being conditionally derived, p_2 gets defeated at line 7 on the condition $\{\neg p_1\}$. This branch of the proof corresponds to the preferred extension $\{a_1, a_4\}$ in the sense that the proof proceeds under the condition that p_1 is valid. The only Ω_P-minimally abnormal $\mathbf{AL_A}$-model validating p_1 is the one also validating p_4 and $\neg p_2, \neg p_3$. Analogously lines 8–9 correspond to the preferred extension $\{p_2, p_4\}$. At line 13 the Ω_P-minimal Dab-formula $\neg p_1 \vee \neg p_2$ is derived. At line 22 another Ω_P-minimal Dab-formula is derived, namely $\neg p_3$.

Note that, unlike the marking procedure for logic $\mathbf{AL_P}$ (see Definition 7.3.4), at line 13 we do not mark lines 6–9. The reason for this is that, concerning the marking conditions for normal selections, we would have to derive $\neg p_1$ (resp. $\neg p_2$) at an unmarked line on a condition $\Delta \subseteq \Omega_{\twoheadrightarrow}$ in order to mark lines derived on the condition $\{\neg p_1\}$ (resp. $\{\neg p_2\}$).

In our example we have two minimal choice sets at line 24, namely $\{\neg p_1, \neg p_3\}$ and $\{\neg p_2, \neg p_3\}$. Therefore we have two Ω_P-minimally abnormal $\mathbf{AL_A}$-models: M_a and M_b where $\mathrm{Ab}^{\mathbf{L_A}}_{\Omega_P}(M_a) = \{\neg p_2, \neg p_3\}$ and $\mathrm{Ab}^{\mathbf{L_A}}_{\Omega_P}(M_b) = \{\neg p_1, \neg p_3\}$. It is easy to see that $M_a \vDash_{\mathbf{L_A}} p_1, p_4, \neg p_2, \neg p_3$ and $M_b \vDash_{\mathbf{L_A}} p_2, p_4, \neg p_1, \neg p_3$. M_a (resp. M_b) therefore corresponds to the preferred extension $\{a_1, a_4\}$ (resp. corresponds to the other preferred extension $\{a_2, a_4\}$).

What is important to notice in this example is that we have

$$\Gamma^4_{\mathsf{A}} \vdash_{\mathbf{ALC_P}} p_1$$
$$\Gamma^4_{\mathsf{A}} \vdash_{\mathbf{ALC_P}} p_2$$
$$\Gamma^4_{\mathsf{A}} \vdash_{\mathbf{ALC_P}} \neg p_1$$
$$\Gamma^4_{\mathsf{A}} \vdash_{\mathbf{ALC_P}} \neg p_2$$

However, we also have for instance

$$\Gamma^4_{\mathsf{A}} \nvdash_{\mathbf{ALC_P}} p_1 \wedge p_2 \tag{7.1}$$
$$\Gamma^4_{\mathsf{A}} \nvdash_{\mathbf{ALC_P}} p_1 \wedge \neg p_1 \tag{7.2}$$

Note that lines 10–12 are marked. While p_1 and p_2 are credulously acceptable in their own respect, as both corresponding arguments —a_1 and a_2— are members of preferred extensions, their conjunction is not. Indeed, there is no preferred extension in which both appear simultaneously. This justifies (7.1). As for (7.2), note that this prevents an explosion. Also, as a_1 is a member of one preferred extension, it is credulously acceptable. Still, a_1 is defeated in another preferred extension and therefore the corresponding p_1 is false in the corresponding model M_b. Of course, we do not want this to lead to the credulous acceptance of '$p_1 \wedge \neg p_1$'.

Let us take a look at the other extension types. Admissible, complete, and

preferred extensions share the same credulously accepted arguments since an argument is credulously accepted with respect to preferred extensions iff it is in a maximal admissible (resp. complete) extension iff it is in an admissible (resp. complete) extension iff it is credulously accepted with respect to admissible (resp. complete) extensions. Thus, logic $\mathbf{ALC_P}$ also models credulous acceptance for admissible (resp. complete) extensions.

Also for grounded extensions we know that an argument is credulously accepted iff it is a member of the unique grounded extension iff it is skeptically accepted. Thus, logic $\mathbf{AL_G}$ also models credulous acceptance.

For semi-stable extensions we proceed in a similar way as for preferred extensions:

$$\mathbf{ALC_S} = \langle \mathbf{L_C}, [\Omega_{\twoheadrightarrow}, \Omega_S, \Omega_P], \\ [\text{simple strategy}, \text{minimal abnormality strategy}, \text{normal selections}] \rangle$$

The semantic consequence relation for $\mathbf{ALC_S}$ is defined similar to Definition 7.5.1 (details can be found in (Straßer and Šešelja 2009)).

Theorem 7.5.1. $\left\{ \begin{array}{cc} (i) & \mathbf{ALC_P} \\ (ii) & \mathbf{ALC_P} \\ (iii) & \mathbf{ALC_S} \end{array} \right\}$ *semantically represents* $\left\{ \begin{array}{l} \textit{max. complete} \\ \textit{preferred} \\ \textit{semi-stable} \end{array} \right\}$ *extensions for argumentation frameworks with at most n arguments.*

Theorem 7.5.2. $\left\{ \begin{array}{cc} (i) & \mathbf{ALC_P} \\ (ii) & \mathbf{ALC_P} \\ (iii) & \mathbf{ALC_P} \\ (iv) & \mathbf{AL_G} \\ (v) & \mathbf{ALC_S} \end{array} \right\}$ *syntactically represents* $\left\{ \begin{array}{l} \textit{admissible} \\ \textit{complete} \\ \textit{preferred} \\ \textit{grounded} \\ \textit{semi-stable} \end{array} \right\}$ *extensions with respect to credulous acceptance for argumentation frameworks with at most n arguments.*

7.6 Discussion

In this discussion section we will localize our results within the context of logical representations of abstract argumentation and highlight some of its advantages.

There are basically two types of logical approaches to argumentation: a meta-level and an object-level approach (see (Caminada and Gabbay 2009)).

The meta-level approaches are often framed in terms of modal logics (see (Grossi 2009, Caminada and Gabbay 2009)). Where $\mathbf{L}$ is such a modal logic, argumentation frameworks are models of $\mathbf{L}$ and arguments are possible worlds. This way extension types and other key properties of argumentation theory can be expressed in terms of the validity of certain formulas in the models.[19]

[19] A different meta-level approach is (Boella et al. 2005). Here, arguments are presented by propositions and extensions are presented by primitives. The authors in (Caminada and

Object-level approaches "model argumentation from within" ((Caminada and Gabbay 2009), p. 134). An AF is represented in terms of premises and arguments as atoms. The logic is supposed to derive acceptable arguments with respect to a given extension type. Obviously, this is the way we motivated our logical framework. Caminada and Gabbay (Caminada and Gabbay 2009) present such a (modal) system for grounded extensions. Our system is clearly more unifying in the sense that it is able to represent all standard extensions of Dung's framework. It is to our knowledge the most unifying object-level logical modeling of argumentation in this sense. This is due to the fact that the adaptive logic framework offers easily adjustable and thus powerful mechanisms that make it possible to obtain a generic proof-theoretic framework for all the different extensions, i.e., with the same representation of AFs as premise sets and only slight variations in the abnormalities and strategies.

It is important to point out that our logical (object-level) approach to abstract argumentation has a variety of advantages that go beyond the capabilities of a simple algorithmic framework that produces skeptically resp. credulously accepted arguments. Some were already mentioned before. For instance, the defeasible character of our modeling allows for the addition of new elements to an AF A^1 on-the-fly, resulting in A^2 (see especially Section 7.3.7 for the technical details). Traditional algorithms have to be applied first to A^1 and then again, from scratch, to A^2. However, our proof theory adjusts to new situations by updating the markings while the argumentation goes on, providing provisional consequences for each step. In this way the dynamics and the rationale of an ongoing argumentation are modeled.

Furthermore, the consequence set of our logics applied to an AF contains more useful information beside the acceptable arguments. Take our Example 7.2.1. As expected, our logic for preferred extensions does derive the only skeptically acceptable p_4. Moreover, the following formulas are derivable: $\varphi_1 = p_1 \vee p_2$, $\varphi_2 = \mathsf{def}\, p_1 \vee \mathsf{def}\, p_2$, $\varphi_3 = \mathsf{def}\, p_3$.[20] The first formula, φ_1, expresses that either a_1 or a_2 is valid in every preferred extension, φ_2 expresses that either a_1 or a_2 is defeated in every preferred extension, and φ_3 expresses that a_3 is defeated in every preferred extension.

Moreover, in some cases the user may take some arguments, X, in A for granted and is thus only interested in, say, preferred extensions that cohere with X, i.e., all preferred extensions E of A such that $E \supseteq X$. In this case the premise set Γ^n_{A} may be enriched by $\{p_i \mid a_i \in X\}$. Our logic for preferred extensions for instance derives inter alia all skeptically acceptable arguments with respect to this subset of all preferred extensions. E.g., for $X = \{a_1\}$ we get the following consequences: $p_4, \mathsf{def}\, p_2, \neg p_2, \mathsf{def}\, p_3, \neg p_3$. This expresses that a_4 is skeptically acceptable in the discussed sense, that a_2 and a_3 are defeated and thus not part of any of the preferred extensions of interest.

Furthermore, we are able to introduce proof theoretic techniques developed

Gabbay 2009) offer, besides the modal systems, also a classical logic meta-level approach (using circumscription).

[20] Note that neither of the following is derivable: $p_1, p_2, \mathsf{def}\, p_1, \mathsf{def}\, p_2$.

for adaptive logics which are interesting for abstract argumentation. For instance, the interpretation of adaptive proofs in terms of argumentation games[21] can be used to model a debate between two parties. Thus, we gain a view on all the standard extension types for abstract argumentation in terms of dialogical games for free on the basis of the presented proof dynamics.

Another advantage of our framework is that it is easily extendable. We give two examples. Often it is not a single argument but rather a bundle of arguments that together attack another argument (see also (Nielsen and Parsons 2006)). Such joint attacks can easily be introduced in our framework by, for instance, allowing for formulas such as $\left(\bigwedge_I p_i\right) \twoheadrightarrow p_j$ expressing that arguments $\{a_i \mid i \in I\}$ attack argument a_j. Our rules (R$\twoheadrightarrow$), (Rad), (R$\bot$), (RCo) and the definition (Def) can be adjusted in a straightforward way.[22] On the other hand, it is interesting to allow for arguments attacking attacks rather than arguments (see e.g., (Modgil 2009)). One such reason is to express a preference for one argument, for instance a_1 attacks $a_2 \rightarrow a_3$ since a_3 is preferred compared to a_2. For our modeling this means that we allow for nested occurrences of the attack operator: $p_1 \twoheadrightarrow (p_2 \twoheadrightarrow p_3)$.[23] It is important to notice that enhancements to abstract argumentation and their combinations can be modeled by minor natural adjustments to our framework and that the modeling is very intuitive.

We have at the end of Section 7.3.2 already pointed out that the external dynamics enabled by the AL representation of abstract argumentation not only opens the possibility to model open, on-going argumentations, but may also be useful for applications in machine learning, belief revision and decision theory.

Moreover, the interpretation of abstract argumentation in terms of adaptive logics offers the possibility to combine the strengths of both frameworks in the modeling of various reasoning forms. Scholars have pointed out the strength of both systems to model for instance defeasible reasoning (see e.g., (Batens et al. 2009, Bondarenko et al. 1997)) or abduction (see e.g., (Neophytos and Antonis 2003, Meheus and Batens 2006)). It remains open for future research to explore these options.

7.7 Conclusion

In this chapter we have presented an adaptive logic characterization of abstract argumentation. Our framework is unifying in the sense that adaptive logic enhancements of one core logic are able to represent all standard extension types for skeptical and credulous acceptance. Skeptically as well as credulously accepted arguments with respect to a given extension type are represented syntactically via the consequence set and semantically in the sense that the

[21] See (Batens 2009) for different variants of argumentation games proposed for adaptive logics, and for instance (Vreeswijk and Prakken 2000) for the interpretation of abstract argumentation in form of games.

[22] This has been realized in (Straßer 2010, Section 8).

[23] We will present the technical details in a future paper.

models correspond to the extensions. The logics differ only insofar as different strategies and different sets of abnormalities are employed.

Moreover, the presented family of logics is apt for the modeling of open-ended argumentations. The logics are able to derive provisional conclusions at different stages of an ongoing discussion. Thus, they explicate the rationale underlying the acceptance of arguments. This is mirrored also by their dynamic proof theory.

It is interesting to notice that the proof dynamics of these logics can be interpreted in terms of argumentation games (see (Batens 2009)). Finally, it should be mentioned that the logics can easily be extended by preferences (Amgoud and Cayrol 1998), values (Bench-Capon 2002), audiences (Bench-Capon et al. 2007) as well as joint attacks (Nielsen and Parsons 2006). We will demonstrate this in a future paper.

Chapter

8 Epilogue

This thesis presented research on certain aspects of theory evaluation, the context of pursuit, and the argumentative approach to methodology and its formal modeling. I would like now to point out some of the questions that can be raised in view of the preceding chapters and that remain open for future research.

First of all, the epistemic goal of robustness, mentioned in Chapter 1 has so far, up to my knowledge, not been used to characterize the scientific knowledge as a whole. Several authors (the most prominent in this respect being William Wimsat (Wimsatt 2007)) have discussed robustness of particular theoretical entities. While the former notion concerns the ability of scientific knowledge to preserve its virtues of explaining and helping us to understand the world by avoiding and in spite of possible scientific crisis and theoretical shifts, the latter notion(s) concern the ability of theories (or smaller theoretical entities) to remain constant throughout scientific development. While the robustness of theories and theoretic entities is supportive of the more general aim of robustness of scientific knowledge, we have argued the latter is not reducible to the former since –in principle– the possibility of crisis cannot be excluded. Hence, the more general notion of robustness plays an important role in the context of theory pursuit. This idea remains to be further developed: on the one hand, by showing how it relates to the other notions of robustness, and on the other hand, by examining a role that different notions of robustness play in different contexts of epistemic justification.

Second, we have presented in Chapter 3 the account of potential coherence by modifying Bonjour's concept of coherence. As we have indicated at the end of this chapter, a similar modification could also be done on the basis of other accounts of epistemic justification. In which way the evaluation in the context of pursuit is effected by changing the basis account, and which of them might be the most suitable ones for this purpose remains to be investigated in future

research.

Next, the distinction between the epistemic and practical pursuit worthiness is closely related to the distinction between epistemic and non-epistemic values. Recent discussions in epistemology and philosophy of science have questioned the viability of this distinction (see, for instance, (Rooney 1992), (Longino 1996), (Douglas 2009)). It remains a task for future research to investigate in which ways these approaches challenge the idea of epistemic justification as such, both in the context of acceptance and in the context of pursuit.

Another issue that has been brought up early in this thesis is the pursuit worthiness of scientifically relevant phenomena, entities, technological developments, etc. Even though we have discussed them as a single set, different from the pursuit and pursuit worthiness of scientific theories, a careful examination of this set should reveal the differences between, for instance, the pursuit of certain phenomena and the pursuit of technological developments. Furthermore, clarifying the difference between the notion of pursuit and the notion of discovery, remains a task for the future research of the context of pursuit.

With respect to our explanatory argumentation framework, its concrete applications to the case studies are yet to be realized. Chapter 6 offered the basic framework that can serve for this task. Moreover, a number of possible enhancements has been indicated. A concrete evaluation of argumentative structures representing rivaling scientific views on the basis of EAFs is another open inquiry. In addition, the question of representing the criteria constituting our account of potential coherence in terms of (possibly enhanced) EAFs is to be tackled as well. Of special interest for this task is the notion of a nested attack (see Section 6.7.3) which allows for attacks to themselves be attacked (without attacking the arguments participating in the attack). For example, the criterion of the programmatic character could be represented in terms of nested attacks: if shortcomings (a) of a given scientific view are represented as an attack on the arguments (b_1–b_n) constituting this view (i.e. $a \rightarrow b_i$ for all $i \leq n$), then such an attack could itself be dismissed if a theory has a programmatic character (c) that addresses this shortcoming ($c \rightarrow (a \rightarrow b_i)$). Such an attack neither attacks a nor b_i. Instead, it dismisses the criticism a of b_i in view of c. For example, in the case of the theory of continental drift, a could be an argument showing that Drift could not prove a mechanism of drifting, while c could be an argument indicating Holmes' hypothesis of convection currents (see Chapters 4 and 6). While c does not attack the fact that there is no sufficient evidence to establish the idea of the mechanism of drift, it attacks its ability to lower the defensive strength of Drift.

Finally, our adaptive logic framework offers a proof-theoretic environment for all the standard extensions of Dung's abstract argumentation. Logics for different selections of arguments belonging to EAFs could be formulated in a similar way, though their concrete realization remains a task for future work. Taking into account the explanatory relations in EAFs these logics may be interesting for the research done on abduction, especially since the abductive reasoning would be approached not from a propositional level, which is usual for the logical systems, but from an argumentative level.

Appendix

Appendix

A Kuhn and Coherentist Epistemology

✎ *This following paper has been published under the same title in Studies in History and Philosophy of Science (Šešelja and Straßer 2009). It is a joint work with Christian Straßer. We are indebted to Erik Weber for comments on an earlier draft of this paper.*

Summary The paper challenges a recent attempt by Jouni-Matti Kuukkanen to show that since Thomas Kuhn's philosophical standpoint can be incorporated into coherentist epistemology, it does not necessarily lead to: (Thesis 1) an abandonment of rationality and rational interparadigm theory comparison, nor to (Thesis 2) an abandonment of convergent realism. Leaving aside the interpretation of Kuhn as a coherentist, we will show that Kuukkanen's first thesis is not sufficiently explicated, while the second one entirely fails. With regard to Thesis 1, we argue that Kuhn's view on inter-paradigm theory comparison allows only for (what we shall dub as) 'the weak notion of rationality', and that Kuukkanen's argument is thus acceptable only in view of such a notion. With regard to Thesis 2, we show that even if we interpret Kuhn as a coherentist, his philosophical standpoint cannot be seen as compatible with convergent realism since Kuhn's argument against it is not 'ultimately empirical', as Kuukkanen takes it to be.

A.1 Introduction

In his recent paper, Jouni-Matti Kuukkanen argues that Thomas Kuhn's philosophical standpoint does not necessarily lead to:

(a) an abandonment of rationality and rational inter-paradigm theory comparison, nor to

(b) an abandonment of scientific realism and the convergence thesis (Kuukkanen 2007, p. 555).

The reason why these two conclusions can be avoided lies, according to Kuukkanen, in the fact that Kuhn's ideas can be interpreted in terms of a coherentist epistemology. More precisely, he argues that a coherentist approach to theory evaluation provides the criteria for a rational inter-paradigm theory comparison (Thesis 1) and is compatible with convergent realism (Thesis 2). The aim of this paper is to discuss Kuukkanen's arguments used for rejecting a) and b). We do not wish to criticize his interpretation of Kuhn as a coherentist, but to show that his theses, built on the basis of this interpretation, are either insufficiently explicated (Thesis 1), or should be entirely rejected (Thesis 2).

With respect to the first point, we will argue that Kuukkanen's notion of rationality, which he uses to characterize Kuhnian inter-paradigm theory comparison, needs to be further elaborated. We will show that Kuhn's ideas on theory choice allow only for (what we shall dub as) "the weak notion of rationality", which is, in fact, compatible with coherentist epistemology, but which might not be acceptable for some philosophers of science. We will conclude that, if Kuukkanen's aim is to interpret Kuhn in terms of some stronger notion of rationality, he is on the wrong track. If, on the other hand, his notion of rationality is the weaker one, introducing coherentist epistemology does not bring any novel insights into Kuhnian inter-paradigm theory comparison, since such a notion of rationality was already explicated by Kuhn himself.

With respect to the second point, we will show that convergent realism cannot be seen as compatible with Kuhn's philosophical standpoint even if we agree that it can be compatible with coherentist epistemology. We will argue that Kuhn's argument against convergent realism is not "ultimately empirical", as Kuukkanen takes it to be, and thus, cannot be refuted on empirical grounds.

The paper is structured as follows. In Section A.2 we present a brief summary of Kuukkanen's arguments. In Section A.3 we discuss the issue of inter-paradigm theory comparison, while Section A.4 elaborates the problem of convergent realism. Section A.5 brings some concluding remarks.

A.2 Kuukkanen on Kuhn

Kuukkanen characterizes Kuhn's work as compatible with coherentist epistemology on the basis of what he calls "Kuhn's epistemological conservatism" and the idea that science is fundamentally problem-solving.[1] Let us have a brief look at each of these points.

First, Kuukkanen argues that the historical perspective underlying Kuhn's approach represents an epistemological framework according to which knowledge is evaluated against a background of accepted beliefs characteristic of a

[1]Since Kuukkanen uses "puzzle solving" and "problem solving" interchangeably, we do the same. For Kuhn's remark on the difference between the two (i.e. between Popperian "problem solving" and his own "puzzle solving") see (Kuhn 1970a, pp. 4–5, 559).

specific scientific paradigm. Further, what needs to be evaluated is not beliefs as such, but the desirability of a particular change of belief. Maintaining that Kuhn adopted a piecemeal approach to theory change in which "it is rational to attempt to improve the justification of the old system, rather than to reject the whole system and try to construct an alternative one", Kuukkanen characterizes Kuhn's position as "epistemological conservatism" (ibid. p. 558).

The second relevant feature of Kuhn's views is that scientific practice is essentially puzzle-solving, and that "the choice between two theories turns, therefore, to the question of whether the suggested alternative manages to solve the puzzle that the older theory could not, or whether it can solve more puzzles than the old one" (ibid., p. 558-559).

In view of these two theses Kuukkanen argues that Kuhn's standpoint can be incorporated into coherentist epistemology. He primarily calls upon L. Bonjour's theory of coherence, summarizing Bonjour's approach in the following three criteria: consistency of the system, the degree of inferential connections it contains, and the number of unexplained anomalous instances it exhibits (ibid., p. 560). According to Kuukkanen, the first point of his interpretation of Kuhn – epistemological conservatism – fits the coherentist idea that the system should not be changed if that results in a decrease of its coherence, and the other way around: if coherence can be increased, the system ought to be changed. The second point – puzzle (or problem) solving – is characterized as a natural component of coherentist epistemology: "Problems, that is, phenomena unexplained by the machinery of the set, decrease the number and strength of inferential relations between the components of the set, making the system less coherent." (ibid., p. 561). In other words, problem solving can be described as coherence-increasing activity.

Although Kuukkanen remarks that Kuhn himself might not have agreed with the description of himself as a coherentist, he points out that there are indications in Kuhn's work that actually go in the direction of coherentism. In this paper we are not going to discuss whether such an interpretation is valid. Kuukkanen himself mentions a number of obstacles for incorporating Kuhn into coherentist epistemology. What we are interested in is the following question: provided we agree with the interpretation of Kuhn as a coherentist, are Kuukkanen's points regarding Kuhn's view on the rationality of inter-paradigm theory comparison and convergent realism acceptable? The following two sections are devoted to these issues.

A.3 Inter-Paradigm Theory Comparison and Theory Choice

The first thesis that Kuukkanen argues for is that Kuhn's philosophy does not necessarily lead to an abandonment of rational inter-paradigm theory comparison, since it can be incorporated into coherentist epistemology. However, he makes no remarks on the notion of rationality that is employed here. In this section we present three possible concepts of rationality with respect to the

determination of theory choice. Next, we discuss which of the three can be taken to describe Kuhn's own approach to theory choice. Finally, we explicate the notion of rationality Kuukkanen needs to accept when arguing that Kuhn's standpoint can be incorporated into coherentist epistemology.

Let us begin by distinguishing two concepts of rationality governing theory choice, with respect to the relation between the criteria of theory choice and the determination of choice. In order to speak at all of a theory choice being rational, we presuppose that the criteria used in this process are, generally speaking, shared by the scientific community, that is, scientists in general agree that arguing according to these standards is rational.

1. *Strong notion of rationality*: the choice of a theory is strictly determined by the criteria shared by the scientific community. That is, an application of the criteria leads to a unique theory choice.

2. *Weak notion of rationality*: the choice of a theory is not strictly determined by the shared criteria. That is, the criteria do not provide a linear preference order on the set of theories in question. With respect to the way the criteria are applied, we can make a further distinction between:

   (a) *Moderately weak notion of rationality*: although the preference order one the criteria, as well as the rules of their application, is fixed, a linear order on theories is not guaranteed: the given criteria together with the rules of their application might be insufficient for deciding between two theories.

   (b) *Very weak notion of rationality*: even though the criteria are shared, the rules of their application as well as the preference order on them are not a priori fixed, but are dependent on the particular context and/or the background knowledge, beliefs, values, etc. of an individual scientist.

Let us now move to Kuhn's views on inter-paradigm theory comparison and theory choice. First of all, Kuhn insisted that the theory comparison is done in view of a set of shared criteria (see, for example, (Kuhn 1977, p. 322), (Kuhn 2000, p. 96)). However,

> There is no neutral algorithm for theory-choice, no systematic decision procedure which, properly applied, must lead each individual in the group to the same decision. (Kuhn 2000, p. 200)

Hence, our *strong notion of rationality* does not fit Kuhn's views. Moreover:

> Individually the criteria are imprecise: individuals may legitimately differ about their application to concrete cases. In addition, when deployed together, they repeatedly prove to conflict with one another: accuracy may, for example, dictate the choice of one theory, scope the choice of its competitor. (Kuhn 1977, p. 322)

> ... for purposes of evaluation, one must embed it [a newly proposed law or theory] in a relevant body of currently accepted beliefs — for example, those governing the instruments with which the relevant observations have been made — and then apply to the whole a set of secondary criteria. Accuracy is one of these, consistency with other accepted beliefs is another, breadth of applicability a third, simplicity a fourth, and there are others besides. All these criteria are equivocal, and they are rarely all satisfied at once. Accuracy is ordinarily approximate, and often unavailable. Consistency is at best local ... Simplicity is in the eye of the beholder. And so on. (Kuhn 2000, p. 114)

Even with respect to puzzle-solving (which Kuukkanen, as we have seen, takes to be Kuhn's key criterion of theory choice) Kuhn writes: "Like any other value, puzzle-solving ability proves equivocal in application." (Kuhn 1996, p. 205). It follows that our *moderately weak notion of rationality* does not capture Kuhn's standpoint either. We are thus left with *the very weak notion of rationality*.

So, for Kuhn there are indeed shared standards governing theory choice, but they alone are not sufficient for explaining it:

> One can explain, as the historian characteristically does, why particular men made particular choices at particular times. But for that purpose one must go beyond the list of shared criteria to characteristics of the individuals who make the choice. (Kuhn 1977, p. 324)

These other criteria can be seen as reasons why the shared standards are applied in different ways. Their different application can be explained by different preference orders on the shared criteria, different ways of evaluating the same (shared) criterion, or different parts of a theoretical framework to which the same (shared) criterion is applied (see ibid., p. 334). Thus, instead of an algorithmic path, discussions among scientists often take the path of persuasion:

> ... the superiority of one theory to another is something that cannot be proved in the debate. Instead, I have insisted, each party must try, by persuasion, to convert the other ... Debates over theory-choice cannot be cast in a form that fully resembles logical or mathematical proof. (Kuhn 1996, pp. 198–199)

Kuhn explains that this does not mean that there are no good reasons for being persuaded, or that these reasons are not ultimately decisive for the group of scientists involved in the discussion, or that they are different from the standard criteria of theory choice, such as accuracy, simplicity, etc. (see also (Kuhn 1970a, p. 238, 241, 260–262)). His point is that the reasons used for persuasion "function as values" (Kuhn 1996, p. 199): just as values can be differently

applied and used to argue in favour of different positions, so can otherwise shared criteria be used to argue in favour of different theories.

Now, if we take into account these arguments, then the only way to incorporate Kuhn's standpoint into coherentist epistemology is by accepting that:

1. the criteria of coherence evaluation do not represent an algorithm that, if properly applied, leads each evaluator to the same result;

2. the way in which the coherentist criteria are applied (for example, their preference order, the scope of their application, etc.) can vary, and is in this way context dependent.

Interestingly enough, such a weak notion of rationality is not incompatible with coherentist accounts themselves. Neither Bonjour nor Thagard offers a preference order on the criteria of coherence evaluation or a strict way in which these criteria are to be applied. On the contrary, both conceptions imply a contextual approach to coherence evaluation. In the case of Bonjour, his Doxastic Presumption, as well as his Observation Requirement, directly implies context dependency of coherence evaluation (see (Bonjour 1985), especially p. 119 and p. 283). In the case of Thagard's account, its contextual character is equally obvious: with regards to the comparative coherence evaluation of phlogiston and oxygen theories, Thagard explicitly points out that his model

> is biased towards the oxygen theory, since it was based on the analysis of Lavoisier's argument. [... It] is not intended to represent the point of view of a phlogiston theorist, a neutral observer, or the entire scientific community. (Thagard 1992, p. 85, 88)

Let us now return to Kuukkanen's interpretation of Kuhn. With regards to Kuhn's point that the criteria of theory choice might be differently applied, he writes:

> Yet, this does not make theory choice arbitrary or irrational. The shared values, however differently shaped, seem to lead to the same theory choice by community members, as 'most members of the group will ultimately find one set of arguments than another decisive' (Kuhn 1996, p. 200).
>
> (Kuukkanen 2007, p. 559); [the reference to Kuhn adapted to our list of references]

This remark does not help in clarifying what Kuukkanen understands by "rational". On the one hand, he is aware of Kuhn's point that the criteria of theory evaluation might be differently applied. On the other hand, he omits to mention that the reason *why* community members tend to make the same choices is not explained by the shared criteria alone. The explanation of their unique choice, according to Kuhn, lies in the process of persuasion that enables the majority to accept specific application of the standards governing theory

evaluation.[2] Which notion of rationality Kuukkanen uses here depends on the way in which the link between the criteria and the theory choice is understood. If his point that shared values "lead to the same theory choice" is supposed to mean that the shared criteria always determine the same result *in spite of* their different application, Kuhn is being interpreted in the sense of our strong notion of rationality. Moreover, Kuukkanen writes:

> I show that problem-solving can be unproblematically connected to a coherentist epistemology. What is more, there are indications in Kuhn's writings that he might have accepted this conclusion. *Surprisingly*, this means that Kuhn *implicitly* agreed that there could be a rational inter-paradigm theory comparison. (ibid., p. 556; italics added).

Again, this passage could be understood as suggesting that the idea of rational inter-paradigm theory comparison is not at all obviously or explicitly present in Kuhn's writings. Since, as we have shown, the "very weak notion of rationality" is indeed elaborated by Kuhn,[3] it could be assumed that Kuukkanen's aim is to interpret Kuhn's standpoint in terms of some stronger notion of rationality. In case this really is Kuukkanen's intention, his interpretation is on the wrong track.

If, on the other hand, Kuukkanen uses the term *rational* in the sense of our notion of very weak rationality, his analysis faces a different sort of criticism. First of all, such an idea of rationality might not seem all that *rational* to some philosophers of science, in particular to those for whom Kuhn's conception of rationality is too weak, that is, to those who, contrary to Kuukkanen, think that Kuhn should not be seen as a rationalist because of his weak notion of rationality (ibid., p. 562). Second, and more importantly, it could be asked, why the recourse to coherentism is needed at all if Kuhnian inter-paradigm theory comparison can already be shown to be rational (in the sense of our very weak notion) on the basis of Kuhn's own writings. If Kuhn already explicated his view on theory choice, what is the benefit of incorporating him into coherentist epistemology, with respect to this question? Such a benefit cannot be found in interpreting Kuhn in terms of some stronger notion of rationality, for such an approach would be inconsistent with Kuhn's own standpoint, as it has been shown in this section. But if we are left with the very weak notion, then coherentism does not offer anything new with respect to the issue discussed, and is thus unexplanatory. Moreover, linking coherentist epistemology with Kuhn

[2]This is clear already from the context in which the part of the sentence quoted by Kuukkanen appears. Let us have a look at the entire sentence: "What one must understand, however, is the manner in which a particular set of shared values interacts with the particular experiences shared by the community of specialists to ensure that most members of the group will ultimately find one set of arguments rather than another decisive. That process is persuasion [...]" (Kuhn 1996, p. 200).

[3]We do not wish to argue that Kuhn gave an elaborated *theory* of rationality governing theory choice, but only that his repeated explication of this problem accords with the very weak notion of rationality defined in this paper.

is not at all straight forward, since, for example, Kuhn maintained that the development of science(s) leads to an increased incoherence among scientific disciplines, as Kuukkanen himself remarks (cp. (Kuhn 2000, pp. 98–99) and (Kuukkanen 2007, p. 564)).

It is important to notice that, speaking in principle, there would be a third option left for Kuukkanen: namely, to show that even though Kuhn's conception of rationality is a weak one, it is not necessarily implied from "the core" of his views. In this case, Kuukkanen's aim would be to link some stronger notion of rationality with a part of Kuhn's views. Nevertheless, he explicitly states that he takes Kuhn's philosophy *as a whole* to be at least consistent with coherentist epistemology:

> I will show below in detail that *Kuhn's philosophy* indeed fits with a coherentist epistemology. (Kuukkanen 2007, p. 558; italics added);
>
> ... I believe this extension of his philosophy does not distort his thinking. (ibid., p. 559)
>
> Now we come to the crucial part. We have to assess how epistemological coherentism meshes with *Kuhn's characterization of science as a whole*, and specifically, how it agrees with the criteria that he suggests are used in theory choice. (ibid., p. 560; italics added)
>
> ... all criteria are linked either directly or indirectly via problem-solving to coherence, which makes *Kuhn's philosophy* consistently coherentist. (ibid., p. 561; italics added)

These passages clearly show that the notion of rationality used by Kuukkanen is not supposed to oppose of Kuhn's views taken as a whole, and therefore should not oppose Kuhn's views on the rationality underlying theory choice either.

Thus, we can conclude that Kuukkanen's argument from coherentism to the rationality of Kuhnian inter-paradigm theory comparison is either invalid or unexplanatory.

A.4 Convergent Realism and Correspondence Theory of Truth

Having argued that Kuhn's position can be interpreted in terms of a coherentist epistemology, Kuukkanen goes on to argue that, since convergent realism is compatible with coherentism, it is therefore compatible with Kuhn's views as well.[4] Kuukkanen agrees that the link from coherentism to realism isn't straight forward, but if we can show continuity, increasing coherence and stability over

[4]Even though Kuukkanen does not offer an explicit definition of convergent realism, it is clear that he refers to a view according to which scientific theories can be seen as converging towards the truth in the sense of the correspondence theory of truth (cp. "[...] the realist typically understands truth as correspondence with reality" (ibid., p. 562) and the rest of Section 4 in (Kuukkanen 2007)).

the long run in the history of science, "an argument for the (approximate) truth of theories has some intuitive appeal" (Kuukkanen 2007, p. 564). According to him, Kuhn's rejection of convergent realism is empirically motivated:

> Although Kuhn had some reservations with the regard to the notion of truth-likeness, he assigned to empirical historical research a central role in deciding the issue of convergence.

And a bit further on:

> Although Kuhn argued that the history of science does not yield support for convergent realism (and for an overall increase of coherence in science), convergent realism is not incompatible with his philosophy because Kuhn's argument is ultimately empirical. (ibid., p. 565)

Let us begin by noting that, in view of the above mentioned quote from Kuukkanen, what he actually claims is the following: had Kuhn found sufficient empirically based support for convergent realism, he would have agreed with it (or at least, such an agreement would be consistent with his own philosophical position). Consequently, if further research reveals some good empirically based arguments for realism, that will be sufficient to refute Kuhn's sceptical view on it.

In addition, in order to claim the compatibility between Kuhn's views and convergent realism, Kuukkanen first had to reassure us that Kuhn's position is compatible with one of the most crucial constituents of convergent realism: the correspondence theory of truth. Thus, at the beginning of his article he argues that Kuhn did not successfully reject the correspondence theory of truth. According to Kuukkanen, Kuhn only showed that there is no direct and unproblematic access to truth, but that did not refute the correspondence theory itself:

> Even if we could not assess a match between a theory and reality, it [Kuhn's attack] does not make the idea that truth consists in a relationship of correspondence between *an independent world* and our beliefs, theories, and so on, meaningless. In other words, the correspondence theory is a theory that offers an interpretation of what truth is without any epistemic concern as to whether we can know the truth. (ibid., p. 556; italics added)

By making this point, Kuukkanen is able to argue in the following way: Kuhn never refuted the correspondence theory itself, so his theory is compatible with it, as well as with convergent realism; if historical arguments show stability, continuity and increasing coherence of scientific theories, it is plausible to accept a convergent realist standpoint, which is thus not necessarily incompatible with Kuhn's approach.

The main problem with this line of reasoning is that, according to Kuhn's central ideas, a valid empirically based argument for convergent realism is, principally speaking, *not possible*. In what follows we will first show that Kuhn himself thought of his argument to be an a priori one. Second, we will present this argument as Kuhn's rejection of the very condition of possibility of the convergent realist view – the correspondence theory of truth.

Let us begin with the first point. Although Kuhn referred to a historical meta-induction, he pointed out that his argument does not rely on it:

> ... my generation of philosophers/historians saw ourselves as building a philosophy on observations of actual scientific behavior. Looking back now, I think that the image of what we were up to is misleading. Given what I shall call the historical perspective, one can reach many of the central conclusions we drew with scarcely a glance at the historical record itself. ... And it is taking longer still to realize that, with that perspective achieved, many of the most central conclusions we drew from the historical record *can be derived instead from first principles. Approaching them in that way reduces their apparent contingency* ... . (Kuhn 2000, pp. 111–112; italics added)

Taking the context into account, it is clear that by first principles Kuhn means the principles that constitute scientific practice *as such*. A rejection of convergent realism could thus rely on what we can conclude from the nature of science, i.e. on its key constituents, without which it would be difficult to conceive science in the sense of the term as we know it. But was such an approach undertaken by Kuhn? By presenting his "tripartite conviction" Kuhn answered this question:

> First, the Archimedean platform outside of history, outside of time and space, is gone beyond recall. Second, in its absence, comparative evaluation is all there is. ... And third, if the notion of truth has a role to play in scientific development, which I shall elsewhere argue that it does, then truth cannot be anything like correspondence to reality. ... I've reached that position from principles that must govern all developmental processes, without, that is, needing to call upon actual examples of scientific behavior. (ibid., p. 115)[5]

Furthermore, the objections on his reference to history and sociology of science were not unknown to Kuhn and he opposed them by emphasizing that

[5] A bit further in the same article, Kuhn compares his arguments against an absolute Archimedean platform and the correspondence theory of truth with the ones he is about to present: "This one, unlike the last, is not necessary or an a priori characteristic, but must be suggested by observations." (Kuhn 2000, p. 116). The comparison thus explicitly shows that Kuhn thought of these arguments as a priori and not based on empirical observations.

"the generalizations which constitute received theories in sociology and psychology (and history?) are weak reeds from which to weave a philosophy of science"(Kuhn 1970a, p. 235) .

The crucial part of Kuhn's a priori reasoning against convergent realism is his argument against the correspondence theory of truth. Let us recall that, according to Kuukkanen, the correspondence theory can be seen as compatible with Kuhn's account since Kuhn never succeeded at rejecting the theory itself. We argue that Kuukkanen does not take into account that Kuhn's attack on the correspondence theory of truth is an attack on one of its main constitutive ideas – the notion of the mind-independent world. We shall present a number of places from Kuhn's work that substantiate our point. We begin with his explicit rejection of the possibility of truth as the correspondence to "the one big mind-independent world", and move towards arguments given in the framework of his so-called "Post-Darwinian Kantianism".

> ... truth cannot be anything like correspondence to reality. I am not suggesting, let me emphasize, that there is a reality which science fails to get at. My point is rather that no sense can be made of the notion of reality as it has ordinarily functioned in philosophy of science. (Kuhn 2000, p. 115)

Kuhn, thus, argues not only that the match between the mind and from it independent reality is not assessable, but that this match is *nonsensical*.

> But the natural sciences, dealing objectively with the real world (as they do), are generally held to be immune. Their truths (and falsities) are thought to transcend the ravages of temporal, cultural, and linguistic change. I am suggesting, of course, that they cannot do so. Neither the descriptive nor the theoretical language of natural science provides the bedrock such transcendence would require. (ibid., p. 75)

The reasons for these claims need to be explicated in view of Kuhn's discussion of the notion of world. First of all, Kuhn emphasizes *the world-constitutive role* of intentionality and mental representations (ibid., p. 103), of a lexicon that is always already in place (ibid., p. 86):

> ... different languages impose different structures on the world. ... where the structure is different, the world is different. (ibid., p. 52)
>
> The world itself must be somehow lexicon-dependent. (ibid., p. 77)

What is thus at stake is the notion of a mind-independent, or in Putnam's terms, "ready-made" world. And for the reasons given above, this term is for Kuhn nonsensical. Nevertheless, he warns his readers that this does not imply that the world is somehow mind-dependent: "the metaphor of a mind-dependent world — like its cousin, the constructed or invented world — proves to be deeply misleading" (ibid., p. 103).

How should the notion of world be treated then? Instead of the strict dichotomy between the mind-independent world and our representations of it, Kuhn proposes "a sort of post-Darwinian Kantianism. Like the Kantian categories, the lexicon supplies preconditions of possible experience" (ibid., p. 104). And as the lexical categories change (ibid.), both in a diachronous and a synchronous manner, "the world ... alters with time and from one community to the next" (ibid., p. 102). Kuhn compares a permanent, fixed, and stable foundation "underlying all these processes of differentiation and change" to "Kant's Ding an sich", which "is ineffable, undescribable, undiscussable" (ibid., p. 104). And what replaces the dichotomy of mind/language/thinking and the one big mind-independent world (ibid., p. 120) is the concept of "niche": "the world is our representation of our niche" (ibid., p. 103).

> Those niches, which both create and are created by the conceptual and instrumental tools with which their inhabitants practice upon them, are as solid, real, resistant to arbitrary change as the external world was once said to be. (ibid., p. 120)

Now, what has become of the notion of truth in Kuhn's post-Darwinian Kantianism?[6] Truth can at best be seen as having "only intra-theoretic applications" (Kuhn 1970a, p. 266):

> Evaluation of a statement's truth values is, in short, an activity that can be conducted only with a lexicon already in place. (Kuhn 2000, p. 77)

By contrast, "[t]he ways of being-in-the-world which a lexicon provides are not candidates for true/false" (ibid., pp. 103–104). None of these "form[s] of life", "practice[s]-in-the-world" give "privileged access to a real, as against an invented, world" (ibid., p. 104). Therefore the speech of theories becoming truer "has a vaguely ungrammatical ring: it is hard to know quite what those who use it have in mind." (ibid., p. 115).[7]

Furthermore, if with Kuhn the sciences form a "complex but unsystematic structure of distinct specialties or species" and therefore have to be "viewed as plural" (ibid., p. 119), and if the niches "do not sum to a single coherent whole

[6] Kuukkanen is not the only one who skips over Kuhn's arguments given in the tradition of Kantian philosophy. Brendan Larvor (cp. (Larvor 2003)), for example, argues that "Kuhn worked into his model of science the historicism found in Koyre and Butterfield" (ibid., p. 386), so that his (Kuhn's) claims "that there is no ahistorical standard of rationality by which past episodes may be judged and that science cannot be shown to be heading towards the Truth – [...] now appear as methodological commitments rather than historico-philosophical theses. Kuhn made waves by dropping an historicist stone into a scientific pond." (ibid., p. 389). However, as our discussion shows, Kuhn's views on these issues cannot be reduced to a mere application of the methodological standards, characteristic for the tradition of historicism in which he stood, to philosophy of science.

[7] Kuhn obviously emphasized his proximity to more "continentally minded" traditions in philosophy not just by his explicit "Kantianism", but also by calling upon key notions such as Heidegger's "being-in-the-world" or late Wittgenstein's "forms of life".

of which we and the practitioners of all the individual scientific specialties are inhabitants" (ibid., p. 120), then "there is no basis for talk of science's gradual elimination of all worlds excepting the single real one." (ibid., p. 86).

What these quotes show is that the problem with the correspondence theory is not only in the *correspondence* itself (as Kuukkanen takes it to be), but also in the notion of world that is supposed to participate in this correspondence. But what would it mean to offer a valid argument against the correspondence theory of truth if not to show that one of its constitutive terms, together with the correspondence itself, is meaningless. Thus, when Kuukkanen claims that Kuhn's argument against the correspondence theory is epistemological, this interpretation is acceptable only if epistemology is taken in view of Kuhn's *transcendental* perspective. Bearing this in mind, we have to reject Kuukkanen's claim that Kuhn "failed to understand the nature of the correspondence theory as a non-epistemic theory", for, as we have seen, such a non-epistemic character of the correspondence theory is for Kuhn plainly nonsensical. Once again:

> There is, I think, no theory-independent way to reconstruct phrases like 'really there'; the notion of a match between the ontology of a theory and its 'real' counterpart in nature now seems to me illusive in principle. Besides, as a historian, I am impressed with the implausibility of the view. (Kuhn 1996, p. 206).

We have thus shown that Kuukkanen's arguments against Kuhn's rejection of the correspondence theory of truth, and for the compatibility of the Kuhnian standpoint with convergent realism – both fail. Showing that Kuhn's position can be incorporated into coherentist epistemology cannot help in bringing him closer to convergent realism since a coherentist approach should either be compatible with Kuhn's a priori argument, or if it is incompatible with it, then so much worse for Kuukkanen's idea of incorporating Kuhn into coherentism.

A.5 Conclusion

J. M. Kuukkanen tried to show that by incorporating Kuhn into coherentist epistemology we can reject the claim that Kuhn's philosophical standpoint abandons a rational inter-paradigm theory comparison, as well as the claim that it is incompatible with convergent realism. In this paper we have argued that there are certain problems with Kuukkanen's arguments. On the one hand, we have shown that Kuhn's views on theory comparison and theory choice allow only for, what we have called, "the very weak notion of rationality", and that Kuhn can be interpreted as a coherentist only in view of this notion. On the other hand, we have shown that Kuhn had an argument against convergent realism, which Kuukkanen did not take into account when claiming that Kuhn's standpoint is compatible with it. In both cases Kuukkanen's point faces the following problem: either coherentist epistemology claims the opposite of Kuhn, and is thus incompatible with Kuhn's ideas, or it is compatible with Kuhn's

ideas, in which case this link offers no new and/or surprising insights into Kuhn's philosophical standpoint, with respect to the issues discussed.[8] We would like to conclude with two remarks regarding our arguments.

With respect to the rationality of Kuhnian inter-paradigm theory comparison, it is important to notice that our distinction between three notions of rationality, though not very refined, is sufficient for our point. That is to say, the distinction is not meant to serve discussions on the issue of rationality in general, since it can indeed be further refined. However, the fact that it is exhaustive is sufficient for our claim that Kuhn's position cannot belong to either of the first two categories.

With regard to Kuhn's a priori argument against the correspondence theory of truth and convergent realism, we would like to remark that in order to challenge Kuukkanen's claim that Kuhn's argument against convergent realism was ultimately empirical, it was sufficient to show that Kuhn, in fact, had an a priori argument. The question as to whether Kuhn's argument is a good one, or whether it is a novel one (or only based on arguments that were already given in the continental philosophical tradition) is irrelevant for our point.

[8]This, however, does not mean that showing the possibility of incorporating Kuhn into coherentist epistemology is of no significance at all, for that is an interesting and valuable insight in itself.

Bibliography

Aliseda, A.: 2006, *Abductive Reasoning*, Springer.

Amgoud, L. and Cayrol, C.: 1998, On the acceptability of arguments in preference-based argumentation., *in* G. F. Cooper and S. Moral (eds), *UAI*, Morgan Kaufmann, pp. 1–7.

Amgoud, L. and Cayrol, C.: 2002, A reasoning model based on the production of acceptable arguments, *Annals of Mathematics and Artificial Intelligence* **34**, 197–215.

Amgoud, L. and Vesic, S.: 2009, On revising argumentation-based decision systems, *in* C. Sossai and G. Chemello (eds), *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Vol. 5590, Springer Berlin Heidelberg, Berlin, Heidelberg, chapter 8, pp. 71–82.

Asquith, P. and Hacking, I. (eds): 1981, *PSA 1978: Proceedings of the 1978 biennial meeting of the Philosophy of Science Association*, Philosophy of Science Association.

Batens, D.: 1999, Inconsistency-adaptive logics, *in* E. Orłowska (ed.), *Logic at Work. Essays Dedicated to the Memory of Helena Rasiowa*, Physica Verlag (Springer), Heidelberg, New York, pp. 445–472.

Batens, D.: 2000, Towards the unification of inconsistency handling mechanisms, *Logic and Logical Philosophy* **8**, 5–31. Appeared 2002.

Batens, D.: 2001, Aspects of the dynamics of discussions and logics handling them, Informal Publication, `http://logica.ugent.be/dirk/SmirR3.pdf`.

Batens, D.: 2007, A universal logic approach to adaptive logics, *Logica Universalis* **1**, 221–242.

Batens, D.: 2009, Towards a dialogic interpretation of dynamic proofs, *in* C. Dégremont, L. Keiff and H. Rückert (eds), *Dialogues, Logics and Other Strange Things. Essays in Honour of Shahid Rahman*, College Publications, London, pp. 27–51.

Batens, D. and Meheus, J.: 2006, Recent results by the inconsistency-adaptive labourers, *in* J.-Y. Béziau and W. A. Carnielli (eds), *Paraconsistent Logic with no Frontiers*, Studies in Logic and Practical Reasoning, North-Holland/Elsevier.

Batens, D., Meheus, J. and Provijn, D.: To appear, An adaptive characterization of signed systems for paraconsistent reasoning.

Batens, D., Mortensen, C., Priest, G. and Van Bendegem, J. P. (eds): 2000, *Frontiers of Paraconsistent Logic*, Research Studies Press, Baldock, UK.

Batens, D., Straßer, C. and Verdée, P.: 2009, On the transparency of defeasible logics: Equivalent premise sets, equivalence of their extensions, and maximality of the lower limit, *Logique et Analyse* **52**(207), 281–304.

Bench-Capon, T. J. M.: 2002, Value based argumentation frameworks, *CoRR* **cs.AI/0207059**. informal publication.

Bench-Capon, T. J. M.: 2003, Persuasion in practical argument using value-based argumentation frameworks, *Journal of Logic and Computation* **13**, 429–448.

Bench-Capon, T. J. M., Doutre, S. and Dunne, P. E.: 2007, Audiences in argumentation frameworks., *Artificial Intelligence* **171**(1), 42–71.

Bermúdez, J. L.: 2005, *Philosophy of Psychology: A Contemporary Introduction*, Routledge.

Béziau, J.-Y. and Carnielli, W. A. (eds): 2006, *Paraconsistent Logic with no Frontiers*, Studies in Logic and Practical Reasoning, North-Holland/Elsevier. In print.

Biagioli, M.: 1993, *Galileo Courtier*, The University of Chicago Press, Chicago and London.

Boella, G., Hulstijn, J. and van der Torre, L. W. N.: 2005, A logic of abstract argumentation, *in* S. Parsons, N. Maudet, P. Moraitis and I. Rahwan (eds), *Argumentation in Multi-Agent Systems*, Vol. 4049 of *Lecture Notes in Computer Science*, Springer, pp. 29–41.

Bondarenko, A., Dung, P. M., Kowalski, R. A. and Toni, F.: 1997, An abstract, argumentation-theoretic approach to default reasoning, *Artificial Intelligence* **93**, 63–101.

Bonjour, L.: 1985, *The Structure of Empirical Knowledge*, Harvard University Press, Cambridge, MA.

Bonjour, L.: 1989, Replies and clarifications, *in* J. W. Bender (ed.), *The Current State of the Coherence Theory*, Kluwer Academic Publishers, pp. 276–292.

Brewka, G. and Lang, J. (eds): 2008, *Principles of Knowledge Representation and Reasoning: Proceedings of the Eleventh International Conference, KR 2008, Sydney, Australia, September 16-19, 2008*, AAAI Press.

Brooks, C.: 1949, *Climate Through The Ages: A Study of the Climatic Factors and Their Variations*, Ernest Benn Limited, London.

Calcott, B.: 2011, Wimsatt and the robustness family: Review of Wimsatt's re-engineering philosophy for limited beings, *Biology and Philosophy* **26**, 281–293.

Caminada, M.: 2006, Semi-stable semantics, *Computational Models of Argument*, IOS Press, pp. 121–132.

Caminada, M. and Gabbay, D.: 2009, A logical account of formal argumentation, *Studia Logica* **93**(2), 109–145.

Carrier, M.: 2010, Knowledge, politics, and commerce: Science under the pressure of practice, *in* M. Carrier and A. Nordsmann (eds), *Science in the Context of Application. Methodological Change, Conceptual Transformation, Cultural Reorientation*, Springer, Dordrecht.

Cayrol, C., de Saint-Cyr, F. D. and Lagasquie-Schiex, M.-C.: 2008, Revision of an argumentation system, *in* (Brewka and Lang 2008), pp. 124–134.

Cayrol, C., Doutre, S. and Mengin, J.: 2003, On decision problems related to the preferred semantics for argumentation frameworks., *Journal of Logic and Computation* **13**(3), 377–403.

Cayrol, C. and Lagasquie-Schiex, M.-C.: 2005, On the acceptability of arguments in bipolar argumentation frameworks, *in* L. Godo (ed.), *ECSQARU*, Vol. 3571 of *Lecture Notes in Computer Science*, Springer, pp. 378–389.

Chang, H.: 2004, *Inventing Temperature: Measurement and Scientific Progress*, Oxford University Press, Oxford, New York.

Coste-Marquis, S., Devred, C., Konieczny, S., Lagasquie-Schiex, M.-C. and Marquis, P.: 2007, On the merging of Dung's argumentation systems, *Artificial Intelligence* **171**, 730–753.

Curd, M. V.: 1980, The logic of discovery: An analysis of three approaches, *in* (Nickles 1980), pp. 201–219.

Dascal, M.: 2000, Epistemology and controversies, *in* T. Y. Cao (ed.), *Philosophy of Science: Volume 10 of Proceedings of the Twentieth World Congress of Philosophy*, Philosophers Index Inc., pp. 159–192.

Doppelt, G.: 1978, Kuhn's epistemological relativism: An interpretation and defense, *Inquiry* **21**(1), 33 86.

Douglas, H. E.: 2009, *Science, Policy, and the Value-Free Ideal*, University of Pittsburgh Press.

du Toit, A. L.: 1937, *Our Wandering Continents: an Hypothesis of Continental Drifting*, Oliver and Boyd, Edinburgh, London.

Dung, P. M.: 1993, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning and logic programming., *International Joint Conference on Artificial Intelligence, Proceedings*, pp. 852–859.

Dung, P. M.: 1995, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, *Artificial Intelligence* **77**, 321–358.

Dung, P. M., Mancarella, P. and Toni, F.: 2002, Argumentation-based proof procedures for credulous and sceptical non-monotonic reasoning, *Lecture Notes in Artificial Intelligence* **2408**, 289–310.

Dung, P. M., Mancarella, P. and Toni, F.: 2007, Computing ideal sceptical argumentation, *Artificial Intelligence* **171**, 642–674.

Dung, P. M. and Son, T. C.: 1996, An argumentation-theoretic approach to reasoning with specificity, *Proceedings of the Fifth International Conference on Principles of Knowledge Representation and Reasoning (KR'96)*, Morgan Kaufmann Publischers, Inc., Cambridge, Massachusetts.

Dunne, P. E. and Bench-Capon, T. J. M.: 2003, Two party immediate response disputes: properties and efficiency, *Artificial Intelligence* **149**(2), 221–250.

Falappa, M. A., Kern-Isberner, G. and Simari, G. R.: 2009, Belief revision and argumentation theory, *Argumentation in Artificial Intelligence*, Springer US, chapter 17, pp. 341–360.

Fitton, J.: 1974, Velikovsky mythistoricus, *Chiron* **I**(1,2), 29–36.

Frankel, H.: 1976, Alfred Wegener and the specialists, *Centaurus* **20**(4), 305–324.

Frankel, H.: 1978, Arthur Holmes and continental drift, *The British Journal for the History of Science* **11**(38), 130–150.

Frankel, H.: 1979, The reception and acceptance of continental drift theory as a rational episode in the history of science, *in* S. H. Mauskopf (ed.), *The Reception of Unconventional Science (AAAS Selected Symposia Series)*, Westview Press, pp. 51–89.

Frankel, H.: 1980, Hess's development of his seafloor spreading hypothesis, *in* (Nickles 1980), pp. 345–366.

Frankel, H.: 1981, The non-Kuhnian nature of the recent revolution in the earth sciences, *in* (Asquith and Hacking 1981), pp. 197–214.

Frankel, H.: 1987, The continental drift debate, *in* H. Engelhardt and A. Caplan (eds), *Scientific Controversies: Case studies in the resolution and closure of disputes in science and technology*, Cambridge University Press, Cambridge, pp. 203–248.

Friedman, M.: 1974, Explanation and scientific understanding, *The Journal of Philosophy* **LXXI**(1), 5–19.

Giere, R. N.: 1988, *Explaining Science: A Cognitive Approach*, The University of Chicago Press, Chicago, London.

Gould, S. J.: 1977, *Ever Since Darwin*, Harvard University, chapter The validation of continental drift, pp. 160–167.

Grandy, R. E.: 2000, On the cognitive analysis of scientific controversies, *in* (Peter Machamer and Baltas 2000), pp. 67–77.

Grant, J.: 1978, Classifications for inconsistent theories, *Notre Dame Journal of Formal Logic* **19**(3), 435–444.

Grant, J. and Hunter, A.: 2006, Measuring inconsistency in knowledgebases, *Journal of Intelligent Information Systems* **27**(2), 159–184.

Grant, J. and Hunter, A.: 2008, Analysing inconsistent first-order knowledgebases, *Artificial Intelligence* **172**(8-9), 1064–1093.

Grossi, D.: 2009, Doing argumentation theory in modal logic, *Technical report*, ILLC Technical Report.

Hansson, S. O.: 2003, Ten philosophical problems in belief revision, *Journal of Logic and Computation* **13**(1), 37–49.

Hempel, C.: 1965, *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*, Free Press, New York.

Hempel, C. and Oppenheim, P.: 1948, Studies in the logic of explanation, *Philosophy of Science* **15**(2).

Holmes, A.: 1931, Radioactivity and earth movements, *Transactions of the Geological Society of Glasgow for 1928-29* **18**, 559–606.

Holmes, A.: 1944, *Principles of Physical Geology*, Thomas Nelson and Sons LTD, London, Edinburgh, Paris, Melbourne, Toronto, New York.

Hoyningen-Huene, P.: 2006, Context of discovery versus context of justification and thomas kuhn, *in* (Schickore and Steinle 2006b), pp. 119–131.

Hughes, W.: 1992, *Critical Thinking*, Broadview Press, Petersborough.

Hunter, A.: 2002, Measuring inconsistency in knowledge via quasi-classical models, *Eighteenth national conference on Artificial intelligence*, American Association for Artificial Intelligence, Menlo Park, CA, USA, pp. 68–73.

Hunter, A. and Konieczny, S.: 2005, Approaches to measuring inconsistent information, *Inconsistency Tolerance. Volume 3300 of Lecture Notes in Computer Science*, Springer, pp. 191–236.

Hunter, A. and Konieczny, S.: 2008, Measuring inconsistency through minimal inconsistent sets, *in* (Brewka and Lang 2008), pp. 358–366.

Janssen, J., Cock, M. D. and Vermeir, D.: 2008, Fuzzy argumentation frameworks, *Proceedings of IPMU 2008 (12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems)*, pp. 513–520.

Kitcher, P.: 1981a, Explanatory unification, *Philosophy of Science* **48**(4), 507–531.

Kitcher, P.: 1981b, Explanatory unification, *Philosophy of Science* **48**, 507–531.

Kitcher, P.: 1989, Explanatory unification and the causal structure of the world, *in* W. S. Philip Kitcher (ed.), *Scientific Explanation*, University of Minnesota Press, Minneapolis, pp. 410–505.

Kitcher, P.: 1990, The division of cognitive labour, *The Journal of Philosophy* **87**(1), 5–22.

Kitcher, P.: 2000, Patterns of scientific controversies, *in* (Peter Machamer and Baltas 2000), pp. 21–39.

Kitcher, P.: 2001, *Science, Truth and Democracy*, Oxford University Press, New York.

Kleiner, S. A.: 2003, Explanatory coherence and empirical adequacy: The problem of abduction, and the justification of evolutionary models, *Biology and Philosophy* **18**, 513–527.

Knight, K.: 2002, Measuring inconsistency, *Journal of Philosophical Logic* pp. 77–98.

Kuhn, T.: 1970a, Logic of discovery or psychology of research, *in* (Lakatos and Musgrave 1970), pp. 1–23.

Kuhn, T.: 1970b, Reflections on my critics, *in* (Lakatos and Musgrave 1970), pp. 231–278.

Kuhn, T.: 1977, *The Essential Tension: selected studies in scientific tradition and change*, University of Chicago press, Chicago.

Kuhn, T.: 1996, *Structure of Scientific Revolutions*, 3 edn, The University of Chicago Press, Chicago.

Kuhn, T.: 2000, *The Road since Structure*, University of Chicago Press, Chicago.

Kukla, A.: 2001, Seti: On the prospects and pursuitworthiness of the search for extraterrestrial intelligence, *Studies in History and Philosophy of Science Part A* **32**(1), 31–67.

Kuukkanen, J.-M.: 2007, Kuhn, the correspondence theory of truth and coherentist epistemology, *Studies in History and Philosophy of Science* **38**, 555–566.

Lakatos, I.: 1978, *The methodology of scientific research programmes*, Cambridge University Press, Cambridge.

Lakatos, I. and Musgrave, A. (eds): 1970, *Criticism and the Growth of Knowledge*, Cambridge University Press, Cambridge, MA.

Larvor, B.: 2003, Why did Kuhn's structure of scientific revolutions cause a fuss?, *Studies in History and Philosophy of Science* **34**, 369–390.

Laudan, L.: 1977, *Progress and its Problems: Towards a Theory of Scientific Growth*, Routledge & Kegan Paul Ltd, London.

Laudan, L.: 1980, Why was the logic of discovery abandoned?, *in* (Nickles 1980), pp. 173–184.

Laudan, L.: 1984, *Science and Values*, University of California Press.

Laudan, R.: 1981, The recent revolution in geology and Kuhn's theory of scientific change, *in* (Asquith and Hacking 1981), pp. 227–239.

Laudan, R.: 1987, The rationality of entertainment and pursuit, *in* J. C. Pitt and M. Pera (eds), *Rational Changes in Science: Essays on Scientific Reasoning*, D. Reidel Publishing Company, Dodrecht, Boston, Lancaster, Tokyo, pp. 203–220.

Laudan, R. and Laudan, L.: 1989, Dominance and the disunity of method: Solving the problems of innovation and consensus, *Philosophy of Science* **56**, 221–237.

Le Grand, H. E.: 1988, *Drifting continents and shifting theories*, Cambridge University Press, Cambridge.

Lewis, C. L. E.: 2002, Arthur Holmes' unifying theory: from radioactivity to continental drift, *Geological Society, London, Special Publications* **192**, 167–183.

Longino, H. E.: 1990, *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*, Princeton University Press, Princeton, New Jersey.

Longino, H. E.: 1996, Cognitive and non-cognitive values in science: Rethinking the dichotomy, *in* L. H. Nelson and J. Nelson (eds), *Feminism, Science, and the Philosophy of Science.*, Kluwer Academic Publishers, Dordrecht, pp. 39–58.

Ma, Y., Qi, G., Xiao, G., Hitzler, P. and Lin, Z.: 2009, An Anytime Algorithm for Computing Inconsistency Measurement, *3rd International Conference on Knowledge Science, Engineering and management 3rd International Conference on Knowledge Science, Engineering and management*, Austria, p. not yet specified.
URL: *http://hal.archives-ouvertes.fr/hal-00422891/en/*

Marvin, U. B.: 2001, Reflections on the American rejection of continental drift, *Metascience* **10**(2), 208–217.

Mayes, G. R.: 2000, Resisting explanation, *Argumentation* **14**, 361–380.

McMullin, E.: 1976, The fertility of theory and the unit for appraisal in science, *in* R. S. Cohen, P. K. Feyerabend and M. W. Wartofsky (eds), *Essays in Memory of Imre Lakatos*, Vol. 39 of *Boston Studies in the Philosophy of Science*, D. Reidel Publishing Company, Dordrecht, pp. 395–432.

McMullin, E.: 1982, Values in science, *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Vol. Volume Two: Symposia and Invited Papers, The University of Chicago Press on behalf of the Philosophy of Science Association, pp. 3–28.

McMullin, E.: 1984, The goals of natural science, *Proceedings and Addresses of the American Philosophical Association* **58**(1), 37–64.

Meheus, J. and Batens, D.: 2006, A formal logic for abductive reasoning, *Logic Journal of the IGPL* **14**, 221–236.

Meheus, J. (ed.): 2002, *Inconsistency in Science*, Kluwer, Dordrecht.

Miller, A. I.: 2002, Inconsistent reasoning toward consistent theories, *in* (Meheus 2002), pp. 35–41.

Modgil, S.: 2006, Hierarchical argumentation, *in* M. Fisher, W. van der Hoek, B. Konev and A. Lisitsa (eds), *JELIA*, Vol. 4160 of *Lecture Notes in Computer Science*, Springer, pp. 319–332.

Modgil, S.: 2009, Reasoning about preference in argumentation frameworks, *Artificial Intelligence* **173**(9-10), 901–934.

Moulin, B., Irandoust, H., Bélanger, M. and Desbordes, G.: 2002, Explanation and argumentation capabilities: Towards the creation of more persuasive agents, *Artificial Intelligence Review* **17**(3), 169–222.

Možina, M., Žabkar, J. and Bratko, I.: 2007, Argument based machine learning, *Artificial Intelligence* **171**(10-15), 922–937.

Neophytos, D. and Antonis, K.: 2003, Argumentation with abduction, *Proceedings of the fourth Panhellenic Symposium of Logic*.

Neurath, O.: 1932/1933 (1983), Protocol statement, *in* R. S. Cohen and M. Neurath (eds), *Philosophical Papers 1913-1946*, Reidel, Dordrecht, pp. 91–99.

Nickles, T.: 2006, Heuristic appraisal: Context of discovery or justification?, *in* (Schickore and Steinle 2006b), pp. 159–182.

Nickles, T. (ed.): 1980, *Scientific Discovery: Case Studies*, D. Reidel Publishing Company., Dordrecht.

Nielsen, S. H. and Parsons, S.: 2006, A generalization of Dung's abstract framework for argumentation: Arguing with sets of attacking arguments, *in* N. Maudet, S. Parsons and I. Rahwan (eds), *Argumentation in Multi-Agent Systems*, Vol. 4766 of *Lecture Notes in Computer Science*, Springer, pp. 54–73.

Oren, N. and Norman, T. J.: 2008, Semantics for evidence-based argumentation, *Proceeding of the 2008 conference on Computational Models of Argument*, IOS Press, Amsterdam, The Netherlands, The Netherlands, pp. 276–284.

Oreskes, N.: 1999, *The Rejection of Continental Drift: Theory and Method in American Earth Science*, Oxford University Press, New York, Oxford.

Oreskes, N.: 2001, Authors response, *Metascience* **10**, 217–222.

Pera, M.: 1994, *The Discourses of Science*, The University of Chicago Press, Chicago, London.

Pera, M.: 2000, Rhetoric and scientific controversies, *in* (Peter Machamer and Baltas 2000), pp. 50–66.

Perelman, C. and Olbrechts-Tyteca, L.: 1969, *The New Rhetoric: A Treatise on Argumentation*, University of Notre Dame Press.

Peter Machamer, M. P. and Baltas, A. (eds): 2000, *Scientific Controversies: Philosophical and Historical Perspectives*, Oxford University Press, New York, Oxford.

Pollock, J.: 1987, Defeasible reasoning, *Cognitive Science* **11**(4), 481–518.

Priest, G.: 2002, Inconsistency and the empirical sciences, *in* (Meheus 2002), pp. 119–128.

Reichenbach, H.: 1938, *Experience and Prediction. An Analysis of the Foundations and the Structure of Knowledge*, University of Chicago Press.

Rooney, P.: 1992, On values in science: Is the epistemic/non-epistemic distinction useful?, *PSA 1992*, Vol. 1, The Philosophy of Science Association, pp. 13–22.

Rueger, A.: 1996, Risk and diversification in theory choice, *Synthese* **109**(2), 263–280.

Schickore, J. and Steinle, F.: 2006a, Introduction: Revisiting the context distinction, in *Revisiting Discovery and Justification: Historical and philosophical perspectives on the context distinction* (Schickore and Steinle 2006b), pp. vii–xix.

Schickore, J. and Steinle, F. (eds): 2006b, *Revisiting Discovery and Justification: Historical and philosophical perspectives on the context distinction*, Springer, Netherlands.

Schurz, G.: 1991, Erklärungsmodelle in der Wissenschaftstheorie und in der Künstlichen Intelligenz, *in* H. Stoyan (ed.), *Proceedings of Erklärung im Gespräch – Erklärung im Mensch-Maschine-Dialog*, Springer, pp. 1–42.

Šešelja, D. and Straßer, C.: 2009, Kuhn and coherentist epistemology, *Studies in History and Philosophy of Science* **40**, 322–327.

Šešelja, D. and Straßer, C.: 2011, Abstract argumentation and explanation applied to scientific debates, *Synthese* **in print**.

Šešelja, D. and Straßer, C.: 201x, Epistemic justification in the context of pursuit: A coherentist approach, *Synthese* **in print**.

Solomon, M.: 2001, *Social Empiricism*, MIT press, Cambridge, Massachusetts.

Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J. and Doyle, J.: 2004, Robustness of cellular functions, *Cell* **118**(6), 675–685.

Stewart, J. A.: 1990, *Drifting Continents & Colliding Paradigms: Perspectives on the Geoscience Revolution*, Indiana University Press, Bloomington.

Straßer, C.: 2010, *Defeasible Reasoning With Applications in Argumentation, Normative Reasoning and Default Reasoning*, PhD thesis, Ghent University.

Straßer, C. and Šešelja, D.: 2009, Towards the proof-theoretic unification of Dung's argumentation framework: An adaptive logic approach — Appendix, `http://logica.ugent.be/centrum/preprints/AFAL-appendix.pdf`.

Straßer, C. and Šešelja, D.: 2011, Towards the Proof-theoretic Unification of Dung's Argumentation Framework: an Adaptive Logic Approach, *Journal of Logic and Computation* **21**(2), 133–156.

Thagard, P.: 1981, The autonomy of a logic of discovery, *in* L. Sumner, J. G. Slater and F. Wilson (eds), *Pragmatism and Purpose, Essays presented to Thomas A. Goudge*, University of Toronto Press, pp. 248–260.

Thagard, P.: 1992, *Conceptual Revolutions*, Princeton University Press, Princeton.

Thagard, P.: 2000, *Coherence in Thought and Action*, MIT Press.

Thagard, P.: 2007, Coherence, truth, and the development of scientific knowledge, *Philosophy of Science* **74**(1), 28–47.

Tursman, R.: 1987, *Peirce's theory of scientific discovery*, Indiana University Press. Bloomington and Indianapolis.

van Helden, A.: 1974, The telescope in the seventeenth century, *Isis* **65**(1), 38–58.

Verheij, B.: 1996, Two approaches to dialectical argumentation: Admissible sets and argumentation stages, *In Proceedings of the biannual International Conference on Formal and Applied Practical Reasoning (FAPR) workshop*, pp. 357–368.

Vreeswijk, G. A. W. and Prakken, H.: 2000, Credulous and sceptical argument games for preferred semantics, *Lecture Notes in Computer Science* **1919**, 239–253.

Watts, A. B.: 2001, *Isostasy and Flexure of the Litosphere*, Cambridge University Press, Cambridge.

Weber, E.: 1999, Unification: What is it, how do we reach and why do we want it?, *Synthese* **118**(3), 479–499.

Wegener, A. L.: 1912, Die Entstehung der Kontinente, *Geologische Rundschau* **3**, 276–292.

Wegener, A. L.: 1915, *Die Entstehung der Kontinente und Ozeane*, Friederich Vieweg und Sohn, Brunswick.

Wegener, A. L.: 1966, *The Origin of Continents and Oceans, (4th ed)*, Methuen, New York, Dover. First published 1929.

Weinert, F.: 2009, *Copernicus, Darwin, and Freud: Revolutions in the History and Philosophy of Science*, Wiley-Blackwell.

Whitt, L. A.: 1990, Theory pursuit: Between discovery and acceptance, *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Vol. 1, pp. 467–483.

Whitt, L. A.: 1992, Indices of theory promise, *Philosophy of Science* **59**, 612–634.

Wimsatt, W. C.: 2007, *Re-engineering philosophy for limited beings: piecewise approximations to reality*, Harvard University Press, Cambridge.